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

Звіт за темою:
«Статистичні критерії на відкритий текст»

Виконав студент
групи ФІ-32мн
Кріпака Ілля

Київ — 2024

1 Мета практикуму

Практично ознайомитися із принципами статистичних методів розрізнення змістовного тексту від випадкової послідовності, порівняти їх.

1.1 Постановка задачі та варіант

Треба виконати	**Зроблено**
Знайти достатню кількість текстів українських творів	✓
Реалізувати усі алгоритми розрізнення змістовного тексту	✓
Показати детермінитичну та стохастичну функції	✓
Проаналізувати отримані результати	✓

2 Хід роботи/Опис труднощів

У ході роботи над даною лабораторною роботою було:

- зібрано багато текстів українською мовою на 12мб даних;
- реалізовано усі критерії за варіантом (1.0-1.3, 3.0, 5.1) + структурний реалізований на алгортимі deflate;
- зібрано усі дані та оформлено у один файл.

Не можна не сказати про труднощі, що виникали під час виконання роботи, а саме:

- на початку практикуму помилково виконав половину критеріїв із не свого варіанту (4.0, 5.0);
- під час виконання 4.0 критерію так і не зміг знайти у ньому помилку, тому навіть не включав його результати до фінальної таблички;
- була спроба зробити ще структурний критерій для алгоритму стиснення lzma, але не вийшло через довгий час роботи;
- не можна не згадати про знаходження текстів (витратив багато часу на набір даних для аналізу).

3 Результати дослідження

У ході роботи було на практиці реалізовано та використано алгортими для розрізнення змістовного тексту. Проаналізовано та створено узагальнену табличку із результатами.

4 Особливості реалізації

Перед тим як говорити про результати роботи критеріїв треба спершу розказати про те як реалізовував їх. Трохи нижче на картинках можна побачити порогові значення 4 та таблиці результатів 5, 6, 7, 8, 9, 10, 11. На таблиці результатів та параметрів можна побачити назви критеріїв та їх параметри для виклику. У наступному списку наведу короткий опис як вони були реалізовані та певні нюанси.

Зауваження. У таблиці є наведені певні розбіжності у кількості текстів, що були проаналізовані. Спершу для тих текстів, що бралися із змістовного тексту було проаналізовано увесь текст, уже не став обрізати можливі входи для критеріїв. Саме тому для $L = 10$ можна побачити 1 млн входів, але на противагу цьому тексти, що треба було генерувати вручну, то там було виставлено константне значення, що було надане у умові лабораторної роботи.

Зауваження. Також у таблиці можна побачити назви, що будуть збивати із пантелику, а саме: 1.3.*own* та 1.1.*custom*. Це ідентичні назви, що позначають те, що цей критерій є модифікованим. Причина чому це залишив, не хотів виправляти усі таблички заново.

1. **Критерій** 1.0 Даний критерій перевіряє певну кількість заборонених l-грам на наявність у певному шматку тексту, що містить (10|100|1000|10000) символів. Параметр h_p визначає кількість заборонених біграм, що будуть використані для перевірки тексту на «справжність». У даному випадку для монограм було вибрано всього 8 штук із 32 символів, а для біграм було вибрано 135 шт. із 1024 шт..

   У результаті можна підсумувати, що критерій вів себе так.

   - Для змістовного тексту в цілому алгоритм себе показав непогано, але у певних випадках навіть у змістовному тексті познаходило заборонені монограми/біграми.
   - Для шифру Віженера із ключем довжини $r = 1$ критерій на малих значеннях (для біграм) показав себе погано, так як багато словосполучень було обрізано, а от для великої довжини, що для монограм, що для баграм він справляється досить добре.
   - Для шифру Віженера із ключем довжини $r = 5$ поведінка у більшості є ідентично до попереднього пункту.
   - Для шифру Віженера із ключем довжини $r = 10$ результат лише змінюється для першого сету n-грам. Там лише для біграм у половині випадків можна назвати текст змістовним.
   - Для шифру афінної та афінної біграмної підстановки $r = 5$ точно так само як попередні критерії для маленьких довжин може давати хибну відповідь.
   - Для випадкового тексту критерій працює добре, налогічно до попередніх випадків із шифром Віженера.
   - Для рекурентного співвідношення аналогічна робота відповідно до попереднього критерію.

   У підсумку можна сказати, що цей критерій реалізує примітивну, але досить успішну логіку, що відсікає тексти із l-грамами, що зовсім не використовуються у змістовному тексті.

2. **Критерій** 1.0 **модифікований** Даний критерій є ідентичним до попереднього, окрім того, що тут змінена перевірка на наявність заборонених монограм/біграм. Дана процедура виконується наступним чином:

- Проходимося по всьому тексту та формуємо геш таблицю із значеннями l-грам, що зустрічаються у змістовному тексті. У даному випадку використовував $l = 2, 3, L$;
- Далі при перевірці L-грами на наявність забороненої l-грами перевірялося на наявність заборонених біграм, триграм та L-грам, тобто як було знайдено будь-яку заборонену l-граму цей текст приймався за випадковий. Сама перевірка на наявність виконується за допомогою перевірки наявності монограми/біграми, що знаходяться у «маленькому» тексті у геш таблиці. Тобто саме ті заборонені l-грами, що взагалі не використовуються у змістовному тексті зміожуть відкинути тексти, що були задані певним перетворенням.

```
if has_prohibited_bigram || has_prohibited_three_gram || has_prohibited_l_gram {
    h_1 += 1;
} else {
    h_0 += 1;
}
```

Рис. 1: Умова при оновленні значень гіпотез.

Параметр $Threshold$ у таблиці параметрів означає мінімальну частоту повторень символу, що повинна бути, для того, щоб вважати знайдені l-грами у тексті, за змістовний текст, а не як заборонену l-граму.

У результаті можна підсумувати, що критерій вів себе так.

- Для змістовного тексту критерій показав себе погано, так як накопичити L-грами досить важко, так як це вимагає великих обчислювальних можливостей.
- Для видозміненого тексту критерій показує себе досить добре, відхиляючи усі можливі тексти.

У підсумку можна сказати, що цей критерій погано себе показує на змістовномму тексті, але безпомилково розрізняє видозмінений текст у якому містяться комбінації символів, що взагалі не використовуються у мові.

3. **Критерій** 1.1 Ідея даного критерію полягає у тому, щоб ввести певне порогове значення разом із яким було б простіше визначати змістовні тексти у яких теж може бути певна кількість заборонених біграм. Параметр k_p позначає поріг після якого L-грама із кількістю заборонених l-грам буде вважатися випадковою, а h_p – кількість заборонених l-грам, що потрібна для перевірки.

   У результаті можна підсумувати, що критерій вів себе наступним чином.

   - Для змістовного тексту критерій показав себе добре, адже у більшості текстів не могло набратися значення, що було більше порогового.
   - Для видозміненого тексту критерій показує себе досить добре, відхиляючи усі можливі тексти.

   У підсумку можна сказати, що цей критерій добре себе показує на змістовному та видозміненому тексті.

4. **Критерій** 1.1 **модифікований** Даний видозмінений алгоритм є подібним до попереднього, лише відрізняється у формі знаходження забороненої l-грами у певному тексті. Цей

алгоритм подібний до того, що використовується у критерії 1.0.*custom*. Тобто текст розподіляється на біграми/триграми/L-грами, знаходиться кількість відповідних входжень заборонених граам у певний текст і рахується сума, яка потім перевіряється на умову. Параметр *Threshold* задає певне порогове значення для таблиці частот, щоб відсіювати дуже малі значення для неї.

```
if amount_of_prohibited_bigram
    + amount_of_prohibited_three_gram
    + amount_of_prohibited_l_gram
    >= prohibited_grams_threshold
{
    h_1 += 1;
} else {
    h_0 += 1;
}
```

Рис. 2: Умова для критерію *criterion*.1.1.*custom*.

У результаті можна підсумувати, що критерій вів себе наступним чином.

- Для змістовного тексту критерій показав себе добре, адже у більшості текстів не могло набратися значення, що було більше порогового.
- Для видозміненого тексту критерій показує себе досить погано, лише на $L = 10000$ результат є допустимим.

У підсумку можна сказати, що цей критерій добре себе показує на змістовному, але погано на видозміненому тексту із маленькою довжиною.

5. **Критерій** 1.2 Даний критерій побудований на частотах символів і має свою суть у тому, що якщо у тексті є аномальна кількість заборонених l-грам, що буде перевищувати частоту l-грами у змістовному тексті, то тоді будемо вважати, що текст є не змістовним. Парметрів даний критерій немає.

   Якщо коротко, то у результаті можна сказати, що критерій лише працює для $l = 2$. Це зв'язано із тим фактом, що частоти монограм, або звичайних букв у змістовному тексті є величезними і частоти, що можна здобути у тексті є дуже малими. Вони аж ніяк не можуть преплюнути попередні виміри, напрочут у біграм це виходить. Тому саме цей метод підходить для використання із дво- і більше грамами.

6. **Критерій** 1.2 **модифікований** Даний критерій побудований подібним чином до 1.0.*custom* із якого була взята перевірка на наявність забороненої l-грами, але основна частина взята із звичайного критерію 1.2. Параметр *Threshold* задає певне порогове значення для таблиці частот, щоб відсіювати дуже малі значення для неї. Умова від критерію 1.1.*custom* не змінилася, але змінилося отримання результатів. Саме тут виконується перевірка для біграм/триграм/L-грам у паралельних потоках, де шукається невідповідність частот у текстах.

   У результаті можна підсумувати, що критерій вів себе наступним чином.

- Для змістовного тексту критерій показав себе посередньо, адже приблизно половина текстів не приймається за змістовний.
- Для видозміненого тексту критерій показує себе добре, адже майже в усіх випадках правильно відрізняє видозмінений текст.

У підсумку можна сказати, що цей критерій показує себе посередньо на змістовному тексті, але добре на видозмінених текстах.

7. **Критерій** 1.3 Даний критерій є таким собі узагальненням попередніх і тепер будемо порівнювати аномальну частоту входження l-грам у текст в сукупності. Тобто будемо враховувати лише частоти l-грам із множини заборонених, яка буде задаватися параметром. Параметромо у даному критерії є h_H – потужність множини заборонених l-грам.

   У результаті можна підсумувати, що критерій вів себе наступним чином.

   - Для змістовного тексту критерій показав себе добре, більшість текстів була розпізнана як змістовна.
   - Для видозміненого тексту критерій показує себе дуже погано, адже у сукупності не може набратися такої кількості частоти, як у звичайному тексті.

   У підсумку можна сказати, що цей критерій показує себе добре на змістовному тексті, але досить посередньо на видозмінених текстах.

8. **Критерій** 1.3 **модифікований** Даний критерій працює так само як і попередній, але як і є у моодифікованих алгоритмів, він використовує інше знаходження заборонених l-грам. У моїй реалізації параметр $Threshold$ задає певне порогове значення для таблиці частот, щоб відсіювати дуже малі значення для неї.

   У підсумку можна сказати, що цей критерій показує себе точно так само як не модифікована версія. Тобто добре на довгому змістовному тексті, але досить посередньо на видозмінених текстах.

9. **Критерій** 3.0 Саме цей критерій уже увібрав у себе ідею обчислення ентропії на символ джерела H_i, що дає обчислювати певну «узагальнену» характеристику невизначеності у тексті. Загалом ентропія у змістовному тексті повинна бути майже такою ж самою як у певній частинці. Параметрами у даному критерії виступає певне порогове значення для різниці ентропій, а саме h_H.

   У результаті можна підсумувати, що критерій вів себе наступним чином.

   - Для змістовного тексту критерій показав себе добре лише для великих текстів, адже на меленьких він не встигає набрати достатні значення для проходження порогу.
   - Для видозміненого тексту критерій показує себе добре лише для великих текстів, аналогічно як попередній пункт.

   У підсумку можна сказати, що цей критерій показує себе добре лише на довгих змістовних або видозмінених текстах.

10. **Критерій** 5.0 **та** 5.1 Ці два критерії треба розглядати у сукупності так як вони представляють із себе один і той же алгоритм, тільки із різних сторін. Головню суттю цих алгоритмів є виділити певні «ящики» і назбирати у них l-грами за які буде відповідати ящик. Тобто ящиками у нас будуть виступати певні найменш або найбльш використовувані l-грами у змістовному. Також будемо вважати ящик пустим тоді і тільки тоді, коли

певні l-грама не зустрілася у тексті. Ящики у критерії 5.0 вибираються як l-грами, що зустрічаються найрідше, а у 5.1 вибираються як щонайчастіше. Щодо параметрів, то у обох алгоритмах використовується параметр $Boxes$, що позначає кількість коробок, які дозволяємо заповнювати та певний поріг k_{empty} – кількість пустих коробок по якому будемо визначати чи належить текст до випадкового чи ні.

У результаті можна підсумувати, що критерій 5.0 веде себе наступним чином.

- Для змістовного тексту критерій показав себе добре лише для великих текстів, але для біграм якось себе дивно веде (у половині текстів показує, що змістовний текст відображається як випадковий).
- Для видозміненого тексту критерій показує себе добре так само лише для текстів із біграмним перетворенням. Із монограмами теж якось не заладилося.

А от у критерія 5.1 спростерігається інша поведінка.

- Для змістовного тексту критерій показав себе посередньо лише для одного із найменших текстів.
- Для видозміненого тексту критерій показує себе добре лише для текстів із біграмних перетвореннями на додачу до деяких із монограмних.

У підсумку можна сказати, що критерій 5.0 на найрідних l-грамах показує себе гірше чим критерій 5.1 на найчастіших l-грамах. Із цього можна вивести висновок, що у випадкових текстах буде якомога менше найбільш частих l-грам.

11. **Структурний критерій для deflate** Почнімо із того, що deflate – є алгортимом стистення без втрат, що використовує у комбінації алгоритми LZ77 та кодування Гаффмана. Так як це є кодуванням, то воно буде звичайно стискати дані, але на цьому факті можна розрізняти чи був стиснений змістовний текст чи випадковий. Будемо обчислювати коефіцієнт компресії для побудови нашого критерію. Ідея була у тому, що змістовні тексти мають у собі більше різних букв, що дасть змогу лише трохи стиснути текст, а щодо незмістовного можна сказати навпаки. Параметром $Threshold$ тут вистуває певне порогове значення стиснення тексту.

```rust
fn struct_deflate(
    l_grams: &Vec<String>,
    threshold: f64,
) -> (u64, u64) {
    let (mut h_0 :u64 , mut h_1 :u64 ) = (0, 0);

    for l_gram :&String  in l_grams {
        let compressed :Vec<u8>  = compress(l_gram.as_bytes());;
        let compression_coef :f64  = compressed.len() as f64 / l_gram.as_bytes().len() as f64;
        if compression_coef > threshold {
            h_1 += 1;
        } else {
            h_0 += 1;
        }
    }
    (h_0, h_1)
}
```

Рис. 3: Умова для обчислення структурного критерію.

У результаті можна підсумувати, що цей критерій веде себе наступним чином.

- Для змістовного тексту критерій показав себе загалом добре у всіх випадках.
- Для видозміненого тексту критерій показує себе погано для маленьких значень, але для великих досить точно класифікує тексти.

4.1 Порогові значення

L	Номер критерію	L=1	L=2
10	1_0	h_p = 8	h_p= 135
	1_0_custom	Threshold = 0	
	1_1	k_p = 2, h_p = 7	k_p = 1, h_p = 200
	1_1_custom	k_p = 9	
	1_2	--	
	1_2_custom	Threshold = 25	
	1_3	K_p = 7	K_p = 200
	1_3_own	Threshold = 25	
	3_0	k_H = 1.7	k_H = 3.23
	5_0	Boxes = 9, k_empty = 8	Boxes = 50, k_empty = 46
	5_1	Boxes = 9, k_empty = 4	Boxes = 50, k_empty = 45
	struct_deflate	Threshold = 1.51	
100	1_0	h_p= 3	h_p= 100
	1_0_custom	Threshold = 0	
	1_1	k_p = 6, h_p = 7	k_p = 1, h_p = 175
	1_1_custom	k_p = 38	
	1_2	--	
	1_2_custom	Threshold = 25	
	1_3	K_p = 7	K_p = 175
	1_3_own	Threshold = 25	
	3_0	k_H = 0.3	k_H = 1.45
	5_0	Boxes = 9, k_empty = 7	Boxes = 100, k_empty = 5
	5_1	Boxes = 9, k_empty = 4	Boxes = 50, k_empty = 40
	struct_deflate	Threshold = 0.65	
1_000	1_0	h_p= 3	h_p= 100
	1_0_custom	Threshold = 0	
	1_1	k_p = 7, h_p = 10	k_p = 2, h_p = 150
	1_1_custom	k_p = 496	
	1_2	--	
	1_2_custom	Threshold = 25	
	1_3	K_p = 8	K_p = 150
	1_3_own	Threshold = 25	
	3_0	k_H = 0.25	k_H = 0.69
	5_0	Boxes = 10, k_empty = 3	Boxes = 100, k_empty = 8
	5_1	Boxes = 9, k_empty = 4	Boxes = 100, k_empty = 40
	struct_deflate	Threshold = 0.37	
10_000	1_0	h_p= 3	h_p= 80
	1_0_custom	Threshold = 0	
	1_1	k_p = 10, h_p = 10	k_p = 10, h_p = 150
	1_1_custom	k_p = 5240	
	1_2	--	
	1_2_custom	Threshold = 25	
	1_3	K_p = 10	K_p = 150
	1_3_own	Threshold = 25	
	3_0	k_H = 0.1	k_H = 0.13
	5_0	Boxes = 10, k_empty = 2	Boxes = 200, k_empty = 188
	5_1	Boxes = 9, k_empty = 4	Boxes = 200, k_empty = 1
	struct_deflate	Threshold = 0.35	

Рис. 4: Порогові значення для критеріїв.

Змістовний текст							
L	Номер критерію	H_0	H_1	(α, β)_1	H_0	H_1	(α, β)_2
10	1_0	632043	413249	(0.39, 0.6)	1043876	1416	(0.001, 0.99)
	1_0_custom	565474	479818	(0.45, 0.54)	565474	479818	(0.45, 0.54)
	1_1	996169	49123	(0.04, 0.95)	1035289	10003	(0.0, 0.99)
	1_1_custom	1045292	0	(0, inf)	1045292	0	(0, inf)
	1_2	1045292	0	(0, inf)	1045271	21	(0.0, 0.99)
	1_2_custom	565474	479818	(0.45, 0.54)	565474	479818	(0.45, 0.54)
	1_3	1045292	0	(0, inf)	1045280	12	(0.0, 0.99)
	1_3_own	604760	440532	(0.42, 0.57)	604760	440532	(0.42, 0.57)
	3_0	771884	273408	(0.26, 0.73)	1008163	37129	(0.03, 0.96)
	5_0	104373	940919	(0.9, 0.09)	0	1045292	(inf, 0)
	5_1	166980	878312	(0.84, 0.15)	78065	967227	(0.92, 0.07)
	struct_deflate	1045292	0	(0, inf)	1045292	0	(0, inf)
100	1_0	43307	61222	(0.58, 0.41)	103969	560	(0.005, 0.99)
	1_0_custom	48414	56115	(0.53, 0.46)	48414	56115	(0.53, 0.46)
	1_1	103045	1484	(0.01, 0.98)	97945	6584	(0.06, 0.93)
	1_1_custom	104529	0	(0, inf)	104529	0	(0, inf)
	1_2	104529	0	(0, inf)	104506	23	(0.0, 0.99)
	1_2_custom	48414	56115	(0.53, 0.46)	48414	56115	(0.53, 0.46)
	1_3	104529	0	(0, inf)	104519	10	(0.0, 0.99)
	1_3_own	83181	21348	(0.2, 0.79)	83181	21348	(0.2, 0.79)
	3_0	74960	29569	(0.28, 0.71)	2846	101683	(0.97, 0.02)
	5_0	90033	14496	(0.13, 0.86)	0	104529	(inf, 0)
	5_1	104529	0	(0, inf)	104520	9	(0.0, 0.99)
	struct_deflate	75448	29081	(0.27, 0.72)	75448	29081	(0.27, 0.72)
1_000	1_0	13	10439	(0.99, 0.001)	9943	509	(0.04, 0.95)
	1_0_custom	2587	7865	(0.75, 0.24)	2587	7865	(0.75, 0.24)
	1_1	4555	5897	(0.56, 0.43)	10449	3	(0.0, 0.99)
	1_1_custom	10452	0	(0, inf)	10452	0	(0. inf)
	1_2	10452	0	(0, inf)	10429	23	(0.0, 0.99)
	1_2_custom	2587	7865	(0.75, 0.24)	2587	7865	(0.75, 0.24)
	1_3	10452	0	(0, inf)	10446	6	(0.0, 0.99)
	1_3_own	10452	0	(0, inf)	10452	0	(0, inf)
	3_0	10452	0	(0, inf)	10452	0	(0, inf)
	5_0	10428	24	(0.0, 0.99)	0	10452	(inf, 0)
	5_1	10452	0	(0, inf)	10452	0	(0, inf)
	struct_deflate	69	10383	(0.99, 0.0)	69	10383	(0.99, 0.0)
10_000	1_0	0	1045	(infinity, 0.0)	850	195	(0.18, 0.81)
	1_0_custom	4	1041	(0.9, 0.0)	4	1041	(0.9, 0.0)
	1_1	59	986	(0.94, 0.05)	1019	26	(0.02, 0.97)
	1_1_custom	1045	0	(0. inf)	1045	0	(0. inf)
	1_2	1045	0	(0, inf)	1021	24	(0.02, 0.97)
	1_2_custom	4	1041	(0.99, 0.0)	4	1041	(0.99, 0.0)
	1_3	1045	0	(0, inf)	1045	0	(0, inf)
	1_3_own	1045	0	(0, inf)	1045	0	(0, inf)
	3_0	1045	0	(0, inf)	988	57	(0.05, 0.94)
	5_0	1045	0	(0, inf)	738	307	(0.29, 0.7)
	5_1	1045	0	(0, inf)	745	300	(0.28, 0.71)
	struct_deflate	1045	0	(0, inf)	1045	0	(0, inf)

Рис. 5: Результати критеріїв для змістовного тексту.

Шифр Віженера із ключем довжини r = 1							
L	Номер критерію	H_0	H_1	(α, β)_1	H_0	H_1	(α, β)_2
10	1_0	82577	962715	(0.92, 0.07)	531866	513426	(0.49, 0.5)
	1_0_custom	0	1045292	(inf, 0)	0	1045292	(inf, 0)
	1_1	139092	906200	(0.86, 0.13)	320937	724355	(0.69, 0.3)
	1_1_custom	960352	84940	(0.08, 0.91)	990164	55128	(0.05, 0.94)
	1_2	1045292	0	(0, inf)	928482	116810	(0.11, 0.88)
	1_2_custom	0	1045292	(inf, 0)	0	1045292	(inf, 0)
	1_3	1045292	0	(0, inf)	965666	79626	(0.07, 0.92)
	1_3_own	847584	197708	(0.18, 0.81)	830387	214905	(0.2, 0.79)
	3_0	771884	273408	(0.26, 0.73)	1008163	37129	(0.03, 0.96)
	5_0	730041	315251	(0.3, 0.69)	25	1045267	(0.99, 0.0)
	5_1	502	1044790	(0.99, 0.0)	9	1045283	(0.99, 0.0)
	struct_deflate	1045292	0	(0, inf)	1045292	0	(0, inf)
100	1_0	136	104393	(0.99, 0.001)	2990	101539	(0.97, 0.02)
	1_0_custom	0	104529	(inf, 0)	0	104529	(inf, 0)
	1_1	71930	32599	(0.31, 0.68)	0	104529	(inf, 0)
	1_1_custom	422	104107	(0.99, 0.0)	52756	51773	(0.49, 0.5)
	1_2	104529	0	(0, inf)	16402	88127	(0.84, 0.15)
	1_2_custom	0	104529	(inf, 0)	0	104529	(inf, 0)
	1_3	104529	0	(0, inf)	104526	3	(0.0, 0.99)
	1_3_own	104526	3	(0.0, 0.99)	104504	25	(0.0, 0.99)
	3_0	74960	29569	(0.28, 0.71)	2846	101683	(0.97, 0.02)
	5_0	104529	0	(0, inf)	0	104529	(inf, 0)
	5_1	102714	1815	(0.01, 0.98)	3762	100767	(0.96, 0.03)
	struct_deflate	67384	37145	(0.35, 0.64)	14592	89937	(0.86, 0.13)
1_000	1_0	0	10452	(inf, 0)	0	10452	(inf, 0)
	1_0_custom	0	10452	(inf, 0)	0	10452	(inf, 0)
	1_1	0	10452	(inf, 0)	0	10452	(inf, 0)
	1_1_custom	614	9838	(0.94, 0.05)	1057	9395	(0.89, 0.1)
	1_2	10452	0	(0, inf)	0	10452	(inf, 0)
	1_2_custom	0	10452	(inf, 0)	0	10452	(inf, 0)
	1_3	10452	0	(0, inf)	10452	0	(0, inf)
	1_3_own	10452	0	(0, inf)	10452	0	(0, inf)
	3_0	10452	0	(0, inf)	10452	0	(0, inf)
	5_0	10451	1	(0.0, 0.99)	0	10452	(inf, 0)
	5_1	10452	0	(0, inf)	0	10452	(inf, 0)
	struct_deflate	55	10397	(0.99, 0.0)	0	10452	(inf, 0)
10_000	1_0	0	1045	(inf, 0)	0	1045	(inf, 0)
	1_0_custom	0	1045	(inf, 0)	0	1045	(inf, 0)
	1_1	59	986	(0.94, 0.05)	0	1045	(inf, 0)
	1_1_custom	1045	0	(0, inf)	0	1045	(inf, 0)
	1_2	1045	0	(0, inf)	0	1045	(inf, 0)
	1_2_custom	0	1045	(inf, 0)	0	1045	(inf, 0)
	1_3	1045	0	(0, inf)	1	1044	(0.99, 0.0)
	1_3_own	0	1045	(inf, 0)	1019	26	(0.02, 0.97)
	3_0	1045	0	(0, inf)	1045	0	(0, inf)
	5_0	1045	0	(0, inf)	1045	0	(0, inf)
	5_1	1045	0	(0, inf)	0	1045	(inf, 0)
	struct_deflate	1045	0	(0, inf)	131	914	(0.87, 0.12)

Рис. 6: Результати критеріїв для шифру Віженера із ключем довжини $r = 1$.

4.2 Таблиці із отриманими результатами

5 Висновки

За допомогою реалізації практикуму «Статистичні критерії на відкритий текст» була можливість побудувати на практиці критерії для перевірки текстів на змістовність тексту та про-

Шифр Віженера із ключем довжини r = 5							
L	Номер критерію	H_0	H_1	(α, β)_1	H_0	H_1	(α, β)_2
10	1_0	37779	1007513	(0.96, 0.03)	529780	515512	(0.49, 0.5)
	1_0_custom	0	1045292	(inf, 0)	0	1045292	(inf, 0)
	1_1	290821	754471	(0.72, 0.27)	156069	889223	(inf, 0)
	1_1_custom	1034578	10714	(0.01, 0.98)	1032648	12644	(0.01, 0.98)
	1_2	1045292	0	(inf, 0)	746233	299059	(0.28, 0.71)
	1_2_custom	0	1045292	(inf, 0)	0	1045292	(inf, 0)
	1_3	1045292	0	(0, inf)	1005786	39506	(0.03, 0.96)
	1_3_own	828958	216334	(0.2, 0.79)	861453	183839	(0.17, 0.82)
	3_0	852603	192689	(0.18, 0.81)	1034937	10355	(0.0, 0.99)
	5_0	850943	194349	(0.18, 0.81)	51	1045241	(0.99, 0.0)
	5_1	8786	1036506	(0.99, 0.0)	50	1045242	(0.99, 0.0)
	struct_deflate	1045292	0	(0, inf)	1045292	0	(inf, 0)
100	1_0	3	104526	(0.99, 0.0)	1710	102819	(0.98, 0.01)
	1_0_custom	0	104529	(inf, 0)	0	104529	(inf, 0)
	1_1	4571	99958	(0.95, 0.04)	0	104529	(inf, 0)
	1_1_custom	14816	89713	(0.85, 0.14)	1130	103399	(0.98, 0.01)
	1_2	104529	0	(0, inf)	11824	92705	(0.88, 0.11)
	1_2_custom	0	104529	(inf, 0)	0	104529	(inf, 0)
	1_3	104529	0	(0, inf)	104529	0	(0, inf)
	1_3_own	104527	2	(0.0, 0.99)	104528	1	(0.0, 0.99)
	3_0	104518	11	(0.0, 0.99)	42270	62259	(0.59, 0.4)
	5_0	104529	0	(0, inf)	0	104529	(inf, 0)
	5_1	104376	153	(0.0, 0.99)	267	104262	(0.99, 0.0)
	struct_deflate	1034	103495	(0.99, 0.0)	270	104259	(0.99, 0.0)
1_000	1_0	0	10452	(inf, 0)	0	10452	(inf, 0)
	1_0_custom	0	10452	(inf, 0)	0	10452	(inf, 0)
	1_1	0	10452	(inf, 0)	0	10452	(inf, 0)
	1_1_custom	44	10408	(0.99, 0.0)	2290	8162	(0.78, 0.21)
	1_2	10452	0	(0, inf)	0	10452	(inf, 0)
	1_2_custom	0	10452	(inf, 0)	0	10452	(inf, 0)
	1_3	10452	0	(0, inf)	10452	0	(0, inf)
	1_3_own	10452	0	(0, inf)	10452	0	(0, inf)
	3_0	454	9998	(0.95, 0.04)	10452	0	(0, inf)
	5_0	10452	0	(0, inf)	0	10452	(inf, 0)
	5_1	10452	0	(0, inf)	5352	5100	(0.48, 0.51)
	struct_deflate	0	10452	(inf, 0)	0	10452	(inf, 0)
10_000	1_0	0	1045	(inf, 0)	0	1045	(inf, 0)
	1_0_custom	0	1045	(inf, 0)	0	1045	(inf, 0)
	1_1	0	1045	(inf, 0)	0	1045	(inf, 0)
	1_1_custom	0	1045	(inf, 0)	1031	14	(0.01, 0.98)
	1_2	1045	0	(0, inf)	0	1045	(inf, 0)
	1_2_custom	0	1045	(inf, 0)	0	1045	(inf, 0)
	1_3	1045	0	(0, inf)	1045	0	(0, inf)
	1_3_own	1045	0	(0, inf)	1045	0	(0, inf)
	3_0	0	1045	(inf, 0)	1045	0	(0, inf)
	5_0	1045	0	(0, inf)	1045	0	(0, inf)
	5_1	1045	0	(0, inf)	94	951	(0.91, 0.08)
	struct_deflate	0	1045	(inf, 0)	0	1045	(inf, 0)

Рис. 7: Результати критеріїв для шифру Віженера із ключем довжини $r = 5$.

аналізувати їх роботу у формі аналітичних таблиць. У результаті наведу критерії, які вважаю є сенс використовувати, а які ні для визначення змістовності тексту.

- Рекомендую до використання

Шифр Віженера із ключем довжини r = 10							
L	Номер критерію	H_0	H_1	(α, β)_1	H_0	H_1	(α, β)_2
10	1_0	52599	992693	(0.94, 0.05)	524178	521114	(0.49, 0.5)
	1_0_custom	0	1045292	(inf,0)	0	1045292	(inf, 0)
	1_1	420047	625245	(0.59, 0.4)	167200	878092	(0.84, 0.15)
	1_1_custom	1024259	21033	(0.02, 0.97)	1018674	26618	(0.02, 0.97)
	1_2	1045292	0	(0, inf)	809782	235510	(0.22, 0.77)
	1_2_custom	0	1045292	(inf,0)	0	1045292	(inf,0)
	1_3	1045292	0	(0, inf)	997227	48065	(0.04, 0.95)
	1_3_own	827772	217520	(0.19, 0.8)	827772	217520	(0.2, 0.79)
	3_0	868798	176494	(0.16, 0.83)	1033695	11597	(0.01, 0.98)
	5_0	767969	277323	(0.26, 0.73)	62	1045230	(0.99, 0.0)
	5_1	1972	1043320	(0.99, 0.0)	21	1045271	(0.99, 0.0)
	struct_deflate	1045292	0	(0, inf)	1045292	0	(0, inf)
100	1_0	10	104519	(0.99, 0.0)	512	104017	(0.99, 0.0)
	1_0_custom	0	104529	(inf,0)	0	104529	(inf, 0)
	1_1	4075	100454	(0.96, 0.03)	0	104529	(inf, 0)
	1_1_custom	852	103677	(0.99, 0.0)	2771	101758	(0.97, 0.02)
	1_2	104529	0	(0, inf)	4199	100330	(0.95, 0.04)
	1_2_custom	0	104529	(inf,0)	0	104529	(inf,0)
	1_3	104529	0	(0, inf)	104528	1	(0.0, 0.99)
	1_3_own	104527	2	(0.99, 0.0)	104528	1	(0.0, 0.99)
	3_0	104366	163	(0.0, 0.99)	56139	48390	(0.46, 0.53)
	5_0	104529	0	(0, inf)	0	104529	(inf,0)
	5_1	104457	72	(0.0, 0.99)	233	104296	(0.99, 0.0)
	struct_deflate	245	104284	(0.99, 0.0)	49	104450	(0.99, 0.0)
1_000	1_0	0	10452	(inf,0)	0	10452	(inf, 0)
	1_0_custom	0	10452	(inf,0)	0	10452	(inf, 0)
	1_1	0	10452	(inf,0)	0	10452	(inf, 0)
	1_1_custom	4538	5914	(0.56, 0.43)	1507	8945	(0.85, 0.14)
	1_2	10452	0	(0, inf)	0	10452	(inf, 0)
	1_2_custom	0	10452	(inf,0)	0	10452	(inf,0)
	1_3	10452	0	(0, inf)	10452	0	(0, inf)
	1_3_own	10452	0	(0, inf)	10452	0	(0, inf)
	3_0	0	10452	(inf,0)	10452	0	(0, inf)
	5_0	10452	0	(0, inf)	0	10452	(inf, 0)
	5_1	10452	0	(0, inf)	1967	8485	(0.81, 0.18)
	struct_deflate	0	10452	(inf,0)	0	10452	(inf,0)
10_000	1_0	0	1045	(inf,0)	0	1045	(inf, 0)
	1_0_custom	0	1045	(inf,0)	0	1045	(inf, 0)
	1_1	0	1045	(inf,0)	0	1045	(inf, 0)
	1_1_custom	1043	2	(0.0, 0.99)	986	59	(0.05, 0.94)
	1_2	1045	0	(0, inf)	0	1045	(inf, 0)
	1_2_custom	0	1045	(inf,0)	0	1045	(inf,0)
	1_3	1045	0	(0, inf)	1045	0	(0, inf)
	1_3_own	1045	0	(0, inf)	1045	0	(0, inf)
	3_0	0	1045	(inf,0)	1045	0	(0, inf)
	5_0	1045	0	(0, inf)	1045	0	(0, inf)
	5_1	1045	0	(0, inf)	754	291	(0.27, 0.72)
	struct_deflate	0	1045	(inf,0)	0	1045	(inf,0)

Рис. 8: Результати критеріїв для шифру Віженера із ключем довжини $r = 10$.

– 1.1 лише на маленьких текстах;
– 3.0 на великих текстах;
– 5.1 на текстах, де є підозра, що перетворення було застосоване біграмне;
– $struct.deflate$ на великих текстах.

Шифр афінної підстановки							
L	Номер критерію	H_0	H_1	(α, β)_1	H_0	H_1	(α, β)_2
10	1_0	18488	1026804	(0.98, 0.01)	376150	669142	(0.64, 0.35)
	1_0_custom	0	1045292	(inf, 0)	0	1045292	(inf, 0)
	1_1	152968	892324	(0.85, 0.14)	176388	868904	(0.83, 0.16)
	1_1_custom	988633	56659	(0.05, 0.94)	1015800	29492	(0.02, 0.97)
	1_2	1045292	0	(0, inf)	845682	199610	(0.19, 0.8)
	1_2_custom	0	1045292	(inf, 0)	0	1045292	(inf, 0)
	1_3	1045292	0	(0, inf)	1033843	11449	(0.01, 0.98)
	1_3_own	842786	202506	296254, 0.8062	825423	219869	(0.21, 0.78)
	3_0	771884	273408	(0.26, 0.73)	1008163	37129	(0.03, 0.96)
	5_0	961571	83721	(0.08, 0.91)	0	1045292	(inf, 0)
	5_1	486	1044806	(0.99, 0.0)	38	1045254	(0.99, 0.0)
	struct_deflate	1045292	0	(0, inf)	1045292	0	(0, inf)
100	1_0	2878	101651	(0.97, 0.02)	5011	99518	(0.95, 0.04)
	1_0_custom	0	104529	(inf, 0)	0	104529	(inf, 0)
	1_1	104529	0	(0, inf)	0	104529	(inf, 0)
	1_1_custom	2317	102212	(0.97, 0.02)	301	104228	(0.99, 0.0)
	1_2	104529	0	(0, inf)	8745	95784	(0.91, 0.08)
	1_2_custom	0	104529	(inf, 0)	0	104529	(inf, 0)
	1_3	104529	0	(0, inf)	104527	2	(0.0, 0.99)
	1_3_own	104529	0	(0, inf)	104511	18	(0.0, 0.99)
	3_0	0	104529	(inf, 0)	2846	101683	(0.97, 0.02)
	5_0	104527	2	(0.99, 0.0)	0	104529	(inf, 0)
	5_1	0	104529	(inf, 0)	0	104529	(inf, 0)
	struct_deflate	62942	41587	(0.39, 0.6)	7879	96650	(0.92, 0.07)
1_000	1_0	0	10452	(inf, 0)	0	10452	(inf, 0)
	1_0_custom	0	10452	(inf, 0)	0	10452	(inf, 0)
	1_1	0	10452	(inf, 0)	0	10452	(inf, 0)
	1_1_custom	0	10452	(inf, 0)	10060	392	(0.03, 0.96)
	1_2	10452	0	(0, inf)	0	10452	(inf, 0)
	1_2_custom	0	10452	(inf, 0)	0	10452	(inf, 0)
	1_3	10452	0	(0, inf)	10452	0	(0, inf)
	1_3_own	10452	0	(0, inf)	10452	0	(0, inf)
	3_0	0	10452	(inf, 0)	10	10442	(0.99, 0.0)
	5_0	10452	0	(0, inf)	0	10452	(inf, 0)
	5_1	10452	0	(0, inf)	57	10395	(0.99, 0.0)
	struct_deflate	9191	1261	(0.12, 0.87)	0	10452	(inf, 0)
10_000	1_0	0	1045	(inf, 0)	0	1045	(inf, 0)
	1_0_custom	0	1045	(inf, 0)	0	1045	(inf, 0)
	1_1	0	1045	(inf, 0)	0	1045	(inf, 0)
	1_1_custom	0	1045	(inf, 0)	0	1045	(inf, 0)
	1_2	1045	0	(0, inf)	0	1045	(inf, 0)
	1_2_custom	0	1045	(inf, 0)	0	1045	(inf, 0)
	1_3	1045	0	(0, inf)	121	924	(0.88, 0.11)
	1_3_own	0	1045	(inf, 0)	1045	0	(0, inf)
	3_0	0	1045	(inf, 0)	1045	0	(0, inf)
	5_0	1045	0	(0, inf)	1045	0	(0, inf)
	5_1	1045	0	(0, inf)	0	1045	(inf, 0)
	struct_deflate	1045	0	(0, inf)	0	1045	(inf, 0)

Рис. 9: Результати критеріїв для шифру афінної та афінної біграмної підстановки.

- Не рекомендую до використання або потребує доопрацювання: 1.0, 1.0.*custom*, 1.1.*custom*, 1.2, 1.2.*custom*, 1.3, 1.3.*custom*, 5.0.

Щоб покращити загальну роботу критеріїв їх можна використовувати комбіновано, до прикладу, для коротких текстів один критеріїв, для довгих інший або змінювати параметри.

Випадковий текст							
L	Номер критерію	H_0	H_1	(α, β)_1	H_0	H_1	(α, β)_2
10	1_0	584	9416	(0.94, 0.05)	5010	4990	(0.49, 0.5)
	1_0_custom	0	10000	(inf, 0)	0	10000	(inf, 0)
	1_1	3621	6379	(0.63, 0.36)	1622	8378	(0.83, 0.16)
	1_1_custom	9763	237	(0.02, 0.97)	9759	241	(0.02, 0.97)
	1_2	10000	0	(0, inf)	7440	2560	(0.25, 0.74)
	1_2_custom	0	10000	(inf, 0)	0	10000	(inf, 0)
	1_3	10000	0	(0, inf)	9614	386	(0.03, 0.96)
	1_3_own	8078	1922	(0.19, 0.8)	8075	1925	(0.19, 0.8)
	3_0	8510	1490	(0.14, 0.85)	9885	115	(0.01, 0.98)
	5_0	7878	2122	(0.21, 0.78)	0	10000	(inf, 0)
	5_1	58	9942	(0.99, 0.0)	0	10000	(inf, 0)
	struct_deflate	10000	0	(0, inf)	10000	0	(0.0, inf)
100	1_0	0	10000	(inf, 0)	55	9945	(0.99, 0.0)
	1_0_custom	0	10000	(inf, 0)	0	10000	(inf, 0)
	1_1	296	9704	(0.97, 0.02)	0	10000	(inf, 0)
	1_1_custom	177	9823	(0.98, 0.01)	178	9822	(0.98, 0.01)
	1_2	10000	0	(0, inf)	406	9594	(0.95, 0.04)
	1_2_custom	0	10000	(inf, 0)	0	10000	(inf, 0)
	1_3	10000	0	(0, inf)	10000	0	(0, inf)
	1_3_own	10000	0	(0, inf)	10000	0	(0, inf)
	3_0	9921	79	(0.0, 0.99)	6683	3317	(0.33, 0.66)
	5_0	10000	0	(0, inf)	0	10000	(inf, 0)
	5_1	9998	2	(0.0, 0.99)	79	9921	(0.99, 0.0)
	struct_deflate	1	9999	(0.0, 0.99)	0	10000	(inf, 0)
1_000	1_0	0	10000	(inf, 0)	0	10000	(inf, 0)
	1_0_custom	0	10000	(inf, 0)	0	10000	(inf, 0)
	1_1	0	10000	(inf, 0)	0	10000	(inf, 0)
	1_1_custom	609	9391	(0.93, 0.06)	601	9399	(0.93, 0.06)
	1_2	10000	0	(0, inf)	0	10000	(inf, 0)
	1_2_custom	0	10000	(inf, 0)	0	10000	(inf, 0)
	1_3	10000	0	(0, inf)	10000	0	(0, inf)
	1_3_own	10000	0	(0, inf)	10000	0	(0, inf)
	3_0	0	10000	(inf, 0)	10000	0	(0, inf)
	5_0	10000	0	(0, inf)	0	10000	(inf, 0)
	5_1	10000	0	(0, inf)	6490	3510	(0.35, 0.64)
	struct_deflate	0	10000	(inf, 0)	0	10000	(inf, 0)
10_000	1_0	0	1000	(inf, 0)	0	1000	(inf, 0)
	1_0_custom	0	1000	(inf, 0)	0	1000	(inf, 0)
	1_1	0	1000	(inf, 0)	0	1000	(inf, 0)
	1_1_custom	144	856	(0.85, 0.14)	126	874	(0.87, 0.12)
	1_2	1000	0	(0, inf)	0	1000	(inf, 0)
	1_2_custom	0	1000	(inf, 0)	0	1000	(inf, 0)
	1_3	0	1000	(inf, 0)	1000	0	(0, inf)
	1_3_own	488	512	(0.51, 0.48)	488	512	(0.51, 0.48)
	3_0	0	1000	(inf, 0)	1000	0	(0, inf)
	5_0	1000	0	(0, inf)	1000	0	(0, inf)
	5_1	1000	0	(0, inf)	988	12	(0.01, 0.98)
	struct_deflate	0	1000	(inf, 0)	0	1000	(inf, 0)

Рис. 10: Результати критеріїв для випадкового незмістовного тексту тексту.

Рекурентне співвідношення							
L	Номер критерію	H_0	H_1	(α, β)_1	H_0	H_1	(α, β)_2
10	1_0	417	9583	(0.95, 0.04)	5091	4909	(0.49, 0.5)
	1_0_custom	0	10000	(inf, 0)	0	10000	(inf, 0)
	1_1	3332	6668	(0.66, 0.33)	1186	8814	(0.88, 0.11)
	1_1_custom	9792	208	(0.02, 0.97)	9771	229	(0.02, 0.97)
	1_2	10000	0	(0, inf)	7342	2658	(0.26, 0.73)
	1_2_custom	0	10000	(inf, 0)	0	10000	(inf, 0)
	1_3	10000	0	(0, inf)	9697	303	(0.03, 0.96)
	1_3_own	7292	2708	(0.27, 0.72)	8199	1801	(0.18, 0.81)
	3_0	8749	1251	(0.12, 0.87)	9883	117	(0.01, 0.98)
	5_0	8958	1042	(0.1, 0.89)	0	10000	(inf, 0)
	5_1	0	10000	(inf, 0)	0	10000	(inf, 0)
	struct_deflate	10000	0	(0, inf)	10000	0	(0, inf)
100	1_0	0	10000	(inf, 0)	40	9960	(0.99, 0.0)
	1_0_custom	0	10000	(inf, 0)	0	10000	(inf, 0)
	1_1	208	9792	(0.97, 0.02)	0	10000	(inf, 0)
	1_1_custom	0	10000	(inf, 0)	155	9845	(0.98, 0.01)
	1_2	10000	0	(0, inf)	308	9692	(0.96, 0.03)
	1_2_custom	0	10000	(inf, 0)	0	10000	(inf, 0)
	1_3	10000	0	(0, inf)	10000	0	(0, inf)
	1_3_own	10000	0	(0, inf)	10000	0	(0, inf)
	3_0	9583	417	(0.04, 0.95)	6994	3006	(0.3, 0.69)
	5_0	10000	0	(0, inf)	0	10000	(inf, 0)
	5_1	10000	0	(0, inf)	78	9922	(0.99, 0.0)
	struct_deflate	0	10000	(inf, 0)	0	10000	(inf, 0)
1_000	1_0	0	10000	(inf, 0)	0	10000	(inf, 0)
	1_0_custom	0	10000	(inf, 0)	0	10000	(inf, 0)
	1_1	0	10000	(inf, 0)	0	10000	(inf, 0)
	1_1_custom	209	9791	(0.97, 0.02)	459	9541	(0.95, 0.04)
	1_2	10000	0	(0, inf)	0	10000	(inf, 0)
	1_2_custom	0	10000	(inf, 0)	0	10000	(inf, 0)
	1_3	10000	0	(0, inf)	10000	0	(0, inf)
	1_3_own	10000	0	(0, inf)	10000	0	(0, inf)
	3_0	0	10000	(inf, 0)	10000	0	(0, inf)
	5_0	10000	0	(0, inf)	0	10000	(inf, 0)
	5_1	10000	0	(0, inf)	7011	2989	(0.29, 0.7)
	struct_deflate	0	10000	(inf, 0)	0	10000	(inf, 0)
10_000	1_0	0	1000	(inf, 0)	0	1000	(inf, 0)
	1_0_custom	0	1000	(inf, 0)	0	1000	(inf, 0)
	1_1	0	1000	(inf, 0)	0	1000	(inf, 0)
	1_1_custom	42	958	(0.95, 0.04)	157	843	(0.84, 0.15)
	1_2	1000	0	(0, inf)	0	1000	(inf, 0)
	1_2_custom	0	1000	(inf, 0)	0	1000	(inf, 0)
	1_3	1000	0	(0, inf)	1000	0	(0, inf)
	1_3_own	1000	0	(0, inf)	1000	0	(0, inf)
	3_0	0	1000	(inf, 0)	435	565	(0.56, 0.43)
	5_0	1000	0	(0, inf)	1000	0	(0, inf)
	5_1	1000	0	(0, inf)	990	10	(0.01, 0.99)
	struct_deflate	0	1000	(inf, 0)	0	1000	(inf, 0)

Рис. 11: Результати структурних критеріїв для стисненого тексту .