Міністерство освіти і науки України

Національний технічний університет України

"Київський політехнічний інститут імені Ігоря Сікорського"

Фізико-технічний інститут

КРИПТОГРАФІЯ

КОМП’ЮТЕРНИЙ ПРАКТИКУМ №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконали:

Студенти 3 курсу

Остапова О. А.

Литвин М. Р.

Київ – 2025

Мета роботи

Засвоєння понять ентропії на символ джерела та його надлишковості, вивчення та порівняння різних моделей джерела відкритого тексту для наближеного визначення ентропії, набуття практичних навичок щодо оцінки ентропії на символ джерела.

Постановка задачі

Задача полягає у проведенні статистичного аналізу тексту російською мовою з метою оцінки його інформаційних характеристик, а саме частот появи символів та біграм, обчислення ентропійних показників H_1 та H_2, а також визначення ентропії у підготовленому тексті з пробілами та без них; додатково необхідно використати програму *CoolPinkProgram* для отримання значень H_{10}, H_{20}, H_{30} і на основі отриманих результатів провести оцінку надлишковості російської мови.

Варіант 10

Порядок виконання роботи

1. Уважно прочитати методичні вказівки до виконання комп'ютерного практикуму.
2. Написати програми для підрахунку частот букв і частот біграм в тексті, а також підрахунку H_1 та H_2 за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення H_1 та H_2 на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення H_1 та H_2 на тому ж тексті, в якому вилучено всі пробіли.
3. За допомогою програми CoolPinkProgram оцінити значення H_{10}, H_{20}, H_{30}.
4. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

Хід роботи

Спочатку реалізуємо код на мові python3 з метою визначити частоту входження символів у текст.

Програма виконує наступні завдання (треба редагувати під код)

- Підготовка тексту: видалення нетекстових символів і пунктуації, перетворення великих літер у малі, заміна «ё»→«е», «ъ»→«ь», нормалізація пробілів.
- Підрахунок кількості входжень кожного символу та обчислення їх частот; формування відсортованої таблиці частот.
- Пошук біграм, що перетинаються та не перетинаються, та обчислення їх частот; формування матриці частот біграм.
- Обчислення ентропій H1 та H2 для тексту з пробілами й без пробілів.
- Збереження результатів у текстовому файлі та табличці

Нажаль, у нас не запрацювала програма Cool pink program, тому ми реалізували алгоритм на пайтоні та вручну зробили цей експеримент.

1)

	Метрика	З пробілами	Без пробілів
0	H1	4.3812	4.4604
1	R1 (для H1)	0.1315	0.1079
2	H2 (з перетином)	3.9901	4.1542
3	R2 (для H2 з перетином)	0.2090	0.1692
4	H2 (БЕЗ перетину)	3.9862	4.1522
5	R2 (для H2 без перетину)	0.2098	0.1696

2)

H(10)

===== Результати експерименту =====

Розподіл кількості спроб (спроба: кількість разів): [(1, 24), (2, 13), (3, 5), (4, 1), (6, 1), (17, 2), (18, 1), (19, 1), (20, 1), (25, 1)]

Ймовірності q_i (спроба: ймовірність):

q_1 : 0.4800

q_2 : 0.2600

q_3 : 0.1000

q_4 : 0.0200

q_6 : 0.0200

q_17: 0.0400

q_18: 0.0200

q_19: 0.0200

q_20: 0.0200

q_25: 0.0200

Оцінка умовної ентропії H(n):

Нижня межа: 1.6180

Верхня межа: 2.2088

Середнє значення: 1.9134

H(20)

===== Результати експерименту =====

Розподіл кількості спроб (спроба: кількість разів): [(1, 23), (2, 10), (3, 6), (4, 2), (6, 1), (8, 1), (12, 1), (14, 1), (17, 1), (18, 1), (19, 1), (23, 1), (27, 1)]

Ймовірності q_i (спроба: ймовірність):

q_1 : 0.4600

q_2 : 0.2000

q_3 : 0.1200

q_4 : 0.0400

q_6 : 0.0200

q_8 : 0.0200

q_12: 0.0200

q_14: 0.0200

q_17: 0.0200

q_18: 0.0200

q_19: 0.0200

q_23: 0.0200

q_27: 0.0200

Оцінка умовної ентропії H(n):

Нижня межа: 1.8049

Верхня межа: 2.5484

Середнє значення: 2.1767

H(30)

===== Результати експерименту =====

Розподіл кількості спроб (спроба: кількість разів): [(1, 26), (2, 15), (3, 6), (4, 1), (12, 1), (24, 1)]

Ймовірності q_i (спроба: ймовірність):

q_1 : 0.5200

q_2 : 0.3000

q_3 : 0.1200

q_4 : 0.0200

q_12: 0.0200

q_24: 0.0200

Оцінка умовної ентропії H(n):

Нижня межа: 1.2147

Верхня межа: 1.7174

Середнє значення: 1.4661

Аналіз:

Порівнюючи значення $H_1 = 4.3812$ та $H_2 = 3.9901$, ми бачимо, що $H_2 < H_1$. Це підтверджує теоретичне положення про те, що врахування статистичних зв'язків між сусідніми символами (біграм) зменшує невизначеність. Знання попередньої літери дозволяє звузити коло ймовірних наступних літер, що призводить до зниження ентропії.

Перейдемо до аналізу експериментальних даних для умовної ентропії H(n). Очікувалося, що зі збільшенням довжини відомого контексту n невизначеність наступного символу буде падати, тобто H(10) > H(20) > H(30).

Отримані результати частково підтверджують цю гіпотезу. Значення H(30) ≈ 1.47 є найнижчим, що логічно, оскільки контекст з 29 символів часто дозволяє майже однозначно передбачити наступну літеру.

Однак, було виявлено аномалію: H(20) ≈ 2.18 виявилося вищим за H(10) ≈ 1.91. Цей результат, що суперечить теорії, швидше за все, є наслідком статистичної похибки через малу кількість експериментів (50 спроб). Цілком імовірно, що під час тестування H(20) випадковим чином були обрані більш неоднозначні фрагменти тексту, де передбачення було об'єктивно складнішим. Для отримання більш гладкого та теоретично коректного результату кількість експериментів слід було б значно збільшити.

Якщо розглянути всю послідовність отриманих значень:

H_1 (4.38) > H_2 (3.99) > H(20) (2.18) > H(10) (1.91) > H(30) (1.47)

(тут я розставив їх за спаданням для наочності).

Стає очевидним головний висновок роботи: чим складніша модель джерела і чим довший контекст вона враховує, тим точніше вона описує реальну структуру мови і тим нижчою виходить оцінка ентропії. Перехід від простих моделей (H_1, H_2) до моделей, що враховують довгі залежності (H(n)), знижує оцінку ентропії більш ніж удвічі (з ~4.4 до ~1.5).

Надлишковість мови показує, наскільки її структура відрізняється від абсолютно випадкової послідовності символів. Згідно з розрахунками, проста модель (H_1) виявляє лише ~12% надлишковості. У той же час, модель, що враховує контекст з 29 символів (H(30)), показує, що надлишковість російської мови становить понад 70%. Це означає, що більше 70% тексту є 'зайвими' з точки зору теорії інформації та визначаються структурними закономірностями мови, що має ключове значення для криптоаналізу.

Висновки:

Підтверджено, що ентропія на символ у природній мові значно нижча за максимальну можливу.

Показано, що ускладнення моделі джерела (від H_1 до H_2) та збільшення довжини контексту (від H_2 до H(n)) призводить до суттєвого зниження оцінки ентропії.

Експериментально встановлено, що ентропія російської мови (з урахуванням контексту до 29 символів) становить приблизно 1.5 біт/символ.

Розраховано, що надлишковість російської мови складає понад 70%, що робить її вразливою для методів статистичного криптоаналізу.

Проаналізовано можливі причини статистичних аномалій в експерименті.