

Міністерство освіти і науки України
Національний технічний університет України
"Київський політехнічний інститут імені Ігоря Сікорського"
Фізико-технічний інститут

КРИПТОГРАФІЯ

КОМП'ЮТЕРНИЙ ПРАКТИКУМ №1

Виконала:
студентка
групи ФБ-13
Теплякова Анна

Експериментальна оцінка ентропії на символ джерела відкритого тексту

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.

Порядок виконання роботи:

0. Уважно прочитати методичні вказівки до виконання комп'ютерного практикуму.

1. Написати програми для підрахунку частот букв і частот біграм в тексті, а також підрахунку $1 H$ та $2 H$ за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення $1 H$ та $2 H$ на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення $1 H$ та $2 H$ на тому ж тексті, в якому вилучено всі пробіли.

2. За допомогою програми CoolPinkProgram оцінити значення $(10) H$, $(20) H$, $(30) H$.

3. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

Обраховані за допомогою написаного коду (lab1.py) значення ентропії:

Для тексту з пробілами:

H_1 для тексту з пробілами: 4.367271375829809

$R = 12.654572483403825 \%$

H_2 для тексту з пробілами (біграми перетинаються): 7.915428210063267

$R = 16.625777036701926 \%$

H_2 для тексту з пробілами (біграми не перетинаються, з кроком 2): 7.914868194713824

$R = 16.625777036701926 \%$

Для тексту з видаленими пробілами:

H_1 без пробілів: 4.468635200737929

H_2 без пробілів (біграми перетинаються): 8.304629606275359

Таблиці кількості та частоти літер та біграм в прикладеному текстовому файлі, бо щось вони занадто довгі для нормального виду протоколу

Перейдемо до роботи з CoolPinkProgram

Округлю значення ентропії

$$2,17 < H^{(10)} < 2,81$$

Лабораторная работа №1

Произвольная часть текста:
ч_тяготения_как_все_организмы_подчиняются_биологическим_законам_так_и_сущес

Использованные буквы:
_

Порядок n-граммы:
5 символов
10 символов
15 символов
20 символов
25 символов
30 символов
35 символов
40 символов
45 символов
50 символов

Введенный символ: и

Символ по счету: 2

Номер эксперимента: 50

Неравенство для энтропии:
 $2,17239886094186 < H < 2,81173259480082$

Двоичная таблица угаданных символов:

00000001000000000000000000000000
01000000000000000000000000000000
0000000000000000000000000000010000
0000000000000000000000000000000100
0001000000000000000000000000000000

Вероятности:

q[1] = 0,5
q[2] = 0,08
q[3] = 0,06
q[4] = 0,02
q[5] = 0
q[6] = 0,02
q[7] = 0
q[8] = 0,04
q[9] = 0,02
q[10] = 0
q[11] = 0,04
q[12] = 0
q[13] = 0
q[14] = 0
q[15] = 0,06
q[16] = 0,04
q[17] = 0
q[18] = 0,04
q[19] = 0,02
q[20] = 0
q[21] = 0
q[22] = 0
q[23] = 0
q[24] = 0
q[25] = 0
q[26] = 0
q[27] = 0
q[28] = 0,02
q[29] = 0,02
q[30] = 0,02
q[31] = 0
q[32] = 0

Поле ввода символов:
и

Продолжить Другой

Строка состояния:
Вы угадали. Для продолжения опыта нажмите "Продолжить", или "Другой" для выбора другого порядка

$$2,241 < H^{(20)} < 2,816$$

Лабораторная работа №1

Произвольная часть текста:
ли_в_виду_какого_то_рода_закон_или_правило_честной_игры_или_порядочного_пов

Использованные буквы:
_

Порядок n-граммы:
5 символов
10 символов
15 символов
20 символов
25 символов
30 символов
35 символов
40 символов
45 символов
50 символов

Введенный символ: _ (пробел)

Символ по счету: 1

Номер эксперимента: 50

Неравенство для энтропии:
 $2,24123117994548 < H < 2,81602202727143$

Двоичная таблица угаданных символов:

10000000000000000000000000000000
00000000000000000100000000000000
10000000000000000000000000000000
01000000000000000000000000000000
00000001000000000000000000000000

Вероятности:

q[1] = 0,44
q[2] = 0,18
q[3] = 0,06
q[4] = 0,02
q[5] = 0
q[6] = 0,02
q[7] = 0
q[8] = 0,02
q[9] = 0,02
q[10] = 0
q[11] = 0
q[12] = 0
q[13] = 0
q[14] = 0,02
q[15] = 0
q[16] = 0
q[17] = 0,02
q[18] = 0
q[19] = 0,02
q[20] = 0,08
q[21] = 0,02
q[22] = 0
q[23] = 0
q[24] = 0
q[25] = 0
q[26] = 0,04
q[27] = 0,02
q[28] = 0
q[29] = 0
q[30] = 0,02
q[31] = 0
q[32] = 0

Поле ввода символов:

Продолжить Другой

Строка состояния:
Вы угадали. Для продолжения опыта нажмите "Продолжить", или "Другой" для выбора другого порядка

The screenshot shows a software interface for a laboratory experiment on entropy. The interface is divided into several sections:

- Top Section:** Contains a text input field with the text "бы_я_знал_как_они_способны_раздражать_человека_я_бы_не_удивлялся_да_и_кто_я" and a list of used letters below it.
- Left Panel:** A list of n-gram orders (5, 10, 15, 20, 25, 30, 35, 40, 45, 50) with "30" selected.
- Center Section:** Includes a "Введенный символ:" field with the letter "з", a "Символ по счету:" field with the value "1", and a "Номер эксперимента:" field with the value "51". Below these is a "Поле ввода символов:" field containing the letter "з". At the bottom of this section are two buttons: "Продолжить" and "Другой".
- Right Panel:** Displays the "Неравенство для энтропии:" as $1.68135993381565 < H < 2.28582744250909$ and a "Двоичная таблица угаданных символов:" which is a 32x32 grid of 0s and 1s.
- Bottom Section:** A "Строка состояния:" showing the message "Вы угадали. Для продолжения опыта нажмите 'Продолжить', или 'Другой' для выбора другого порядка".
- Far Right Panel:** A vertical list of probabilities $q[1]$ through $q[32]$, with most values being 0 and a few non-zero values for $q[1]$ through $q[19]$.

$$R = 1 - \frac{H_0}{H_\infty}$$

$$H_0 = \log_2 32 = 5$$

$$H^{(10)}$$

$$\mathbf{H}^{(30)}$$

$$1 - \frac{1,681}{5} < R < 1 - \frac{2,286}{5}$$

$$0,6638 < R < 0,5428$$

Висновки:

Під час виконання роботи, я навчилася практичним шляхом підраховувати частоти букв та біграм на обраному тексті, обраховувати значення ентропії на символ джерела та надлишковість джерела відкритого тексту. За думкою Інтернету, надлишковість російської мови в цілому = 72,6%, а літературної мови = 76,2%, нехай середня надлишковість буде 74%. Моє найвище значення нижче на майже 10 відсотків. Це можна пояснити тим, що в усіх трьох експериментах я вгадувала літери з першого разу найбільше за інші літери (хоча я майже через раз просто починала клацати вздовж по клавіатурі). Бо чим непередбачуваніше повідомлення, тим більша його ентропія.