

МІНІСТЕРСТВО ОСВІТИ І НАУКИ, МОЛОДІ ТА СПОРТУ УКРАЇНИ  
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ  
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»  
ФІЗИКО-ТЕХНІЧНИЙ ІНСТИТУТ

КРИПТОГРАФІЯ

КОМП'ЮТЕРНИЙ ПРАКТИКУМ №1

«Експериментальна оцінка ентропії на символ джерела відкритого тексту»

Виконали:

Студенти групи 3 курсу  
груп ФБ-96 та ФБ-94  
Ігнатенко Артем  
Васюченко Георгій

Київ 2021

Мета:

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:

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 Н та 2 Н за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення 1 Н та 2 Н на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де ймовірності замінити відповідними частотами. Також одержати значення 1 Н та 2 Н на тому ж тексті, в якому вилучено всі пробіли.
2. За допомогою програми CoolPinkProgram оцінити значення (10) Н , (20) Н , (30) Н .
3. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

Хід роботи:

Використані формули:

Ентропія

$$H(Z) = - \sum_{i=1}^n p_i \log p_i .$$

```
def H(freq, n):  
    temp = []  
    for f in freq.values():  
        temp.append(f * math.log(f, 2))  
    temp = sorted(temp)  
    H = -sum(temp) / n  
    return H
```

Надлишковість

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

$$H_0 = \log_2 m ,$$

```
def R(value):  
    R_ = 1 - value / math.log(33, 2)  
    print(R_)  
    return (R_)
```

Задача 1:

Частота монограм без пробілів				Частота монограм з пробілами			
1	frequency_without_spaces			1	frequency_with_spaces		
2	о	0,11459		2		0,173396	
3	е	0,08699		3	о	0,094724	
4	а	0,07957		4	е	0,071907	
5	н	0,06501		5	а	0,065772	
6	и	0,06477		6	н	0,053739	
7	т	0,06468		7	и	0,053542	
8	с	0,05287		8	т	0,053463	
9	в	0,04621		9	с	0,043703	
10	л	0,04591		10	в	0,038196	
11	р	0,04178		11	л	0,037949	
12	к	0,03299		12	р	0,034539	
13	д	0,03198		13	к	0,027269	
14	м	0,0314		14	д	0,026436	
15	у	0,02962		15	м	0,025959	
16	п	0,02741		16	у	0,024483	
17	ь	0,02296		17	п	0,022658	
18	я	0,02134		18	ь	0,018978	
19	ч	0,01808		19	я	0,01764	
20	б	0,01737		20	ч	0,014947	
21	г	0,01687		21	б	0,014357	
22	ы	0,01649		22	г	0,013946	
23	з	0,01538		23	ы	0,013635	
24	ж	0,0114		24	з	0,012713	
25	й	0,01		25	ж	0,00942	
26	х	0,0085		26	й	0,008268	
27	ш	0,00822		27	х	0,007025	
28	ю	0,00561		28	ш	0,006796	
29	э	0,00352		29	ю	0,004638	
30	щ	0,00299		30	э	0,002912	
31	ц	0,00277		31	щ	0,002469	
32	ф	0,00124		32	ц	0,002289	
33	ъ	0,00024		33	ф	0,001027	
				34	ъ	0,0002	





Задача 2:

H(10)

Лабораторная работа №1

Произвольная часть текста:  
ит\_смешно

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:

Символ по счету:

Номер эксперимента: 51

Поле ввода символов:

Продолжить

Другой

Неравенство для энтропии:  
2,46802172251638< H < 3,05585768303304

Двоичная таблица угаданных символов:  
000000100000000000000000000000  
000000100000000000000000000000  
000000100000000000000000000000  
100000000000000000000000000000  
100000000000000000000000000000

Вероятности:

q[1] = 0,42  
q[2] = 0,06  
q[3] = 0,04  
q[4] = 0,02  
q[5] = 0,02  
q[6] = 0,04  
q[7] = 0,14  
q[8] = 0,04  
q[9] = 0,04  
q[10] = 0  
q[11] = 0  
q[12] = 0  
q[13] = 0,06  
q[14] = 0,02  
q[15] = 0  
q[16] = 0,02  
q[17] = 0  
q[18] = 0  
q[19] = 0  
q[20] = 0,02  
q[21] = 0  
q[22] = 0  
q[23] = 0  
q[24] = 0  
q[25] = 0  
q[26] = 0,02  
q[27] = 0  
q[28] = 0  
q[29] = 0  
q[30] = 0  
q[31] = 0,02  
q[32] = 0,02

Строка состояния:

H(20)

Лабораторная работа №1

Произвольная часть текста:  
ным\_то\_есть\_законом

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:

Символ по счету:

Номер эксперимента: 51

Поле ввода символов:

Продолжить

Другой

Неравенство для энтропии:  
2,93334045007081< H < 3,40353232348999

Двоичная таблица угаданных символов:  
000000000000000000000000000001  
100000000000000000000000000000  
010000000000000000000000000000  
100000000000000000000000000000  
010000000000000000000000000000

Вероятности:

q[1] = 0,34  
q[2] = 0,18  
q[3] = 0,04  
q[4] = 0  
q[5] = 0,02  
q[6] = 0  
q[7] = 0  
q[8] = 0  
q[9] = 0,04  
q[10] = 0  
q[11] = 0  
q[12] = 0  
q[13] = 0,02  
q[14] = 0,02  
q[15] = 0,02  
q[16] = 0,02  
q[17] = 0  
q[18] = 0  
q[19] = 0,08  
q[20] = 0,02  
q[21] = 0,02  
q[22] = 0,02  
q[23] = 0,02  
q[24] = 0,02  
q[25] = 0,02  
q[26] = 0  
q[27] = 0,02  
q[28] = 0  
q[29] = 0,04  
q[30] = 0  
q[31] = 0,02  
q[32] = 0,02

Строка состояния:

## H(30)

### Задача 3:

$2.468021 < H(10) < 3.055857$	$0.510739 > R(10) > 0.394207$
$2.933340 < H(20) < 3.403532$	$0.418495 > R(20) > 0.325284$
$1.535043 < H(30) < 2.131663$	$0.695693 > R(30) > 0.577419$

## Висновки:

Здобуто навички вимірювання частот повторювання різних символів в довільному тексті, оцінки ентропії, надлишковості на прикладі російської мови, підрахування частот монограм та біграм у довільному тексті.