

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  
групи ФБ-21  
КАЮН Вероніка  
РУДЮК Олександр

**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

1. 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.

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

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

### Хід роботи

#### Довільний текст

Спочатку підрахуємо частоту букв та біграм у нашому тексті **text2** – текст з пробілами.

Частота букв	Частота біграм з перетином	Частота біграм без перетину
--------------	-------------------------------	--------------------------------

Буква	Частота	Біграма	Частота	Біграма	Частота
	0,123861	вв	5,38E-05	вв	5,57E-05
о	0,093987	ве	0,002961	ед	0,002959
и	0,088443	ед	0,002961	ен	0,011768
е	0,070679	де	0,00323	ие	0,005799
н	0,063358	ен	0,011789	т	0,008594
т	0,059321	ни	0,012865	ех	0,004236
а	0,057921	ие	0,00576	но	0,015329
с	0,048501	е	0,015342	ло	0,006319
р	0,040911	т	0,008559	ги	0,00601
в	0,039834	те	0,009313	и	0,024799
л	0,034397	ех	0,004199	за	0,003137
м	0,028153	хн	0,00393	ни	0,012826
к	0,025785	но	0,015288	ма	0,002536
ь	0,022662	ол	0,009366	ют	0,003653
п	0,021747	ло	0,006298	ц	0,001682
д	0,021155	ог	0,008398	тр	0,004982
я	0,018087	ги	0,005975	ал	0,004826
у	0,017441	ии	0,004952	ьн	0,00317
з	0,016956	и	0,024816	ое	0,001188
г	0,01448	з	0,00183	м	0,005621
б	0,012704	за	0,003122	ес	0,00721
ь	0,012327	ан	0,008182	то	0,006916
х	0,012112	им	0,003337	в	0,009578
ч	0,009366	ма	0,00253	р	0,005494
й	0,008236	аю	0,002153	аз	0,002625
ц	0,007913	ют	0,00366	ви	0,003014

Всі значення можна переглянути у файлі **frequency\_data.xlsx**

Далі обчислюємо ентропію та надлишковість тексту.

```

Ентропія H1 (монограми): 4.61260
Надлишковість R1 (монограми): 0.23961
Ентропія H2 (біграми з перетином): 8.10212
Надлишковість R2 (біграми з перетином): 0.33218
Ентропія H2 (біграми без перетину): 8.10161
Надлишковість R2 (біграми без перетину): 0.33222

```

### Текст без пробілів

Видаляємо пробіли із тексту та зберігаємо у **text3**.

Частота букв	Частота біграм з перетином	Частота біграм без перетину
--------------	-------------------------------	--------------------------------

Буква	Частота	Біграма	Частота	Біграма	Частота
о	0,104827	вв	0,00024	вв	0,00012
и	0,098643	ве	0,003302	ед	0,003722
е	0,078831	ед	0,003842	ен	0,013329
н	0,070665	де	0,003602	ие	0,006004
т	0,066162	ен	0,014289	т	0,002642
а	0,064601	ни	0,014469	ех	0,005043
с	0,054095	ие	0,006484	но	0,016691
р	0,045629	е	0,001081	ло	0,007085
в	0,044429	т	0,002342	ги	0,006724
л	0,038365	те	0,010387	из	0,005163
м	0,0314	ех	0,004743	ан	0,008525
к	0,028758	хн	0,004503	им	0,004323
ы	0,025276	но	0,017051	аю	0,003002
п	0,024256	ол	0,010507	тц	0,00012
д	0,023595	ло	0,007025	тр	0,005283
	0,022813	ог	0,009486	ал	0,006364
я	0,020173	ги	0,006784	ьн	0,004923
у	0,019452	ии	0,008706	ое	0,001561
з	0,018912	из	0,005403	ме	0,004923
г	0,01615	за	0,003482	ст	0,020653
б	0,014169	ан	0,009246	ов	0,013809
ь	0,013749	им	0,004983	ра	0,012248
х	0,013509	ма	0,002942	зв	0,002642
ч	0,010447	аю	0,002402	ит	0,005043
й	0,009186	ют	0,004203	ии	0,008405
ц	0,008826	тц	6E-05	со	0,006004

Всі значення можна переглянути у файлі **frequency\_data\_spaces\_del.xlsx**

```

Ентропія H1 (монограми): 4.63480
Надлишковість R1 (монограми): 0.23321
Ентропія H2 (біграми з перетином): 8.26551
Надлишковість R2 (біграми з перетином): 0.31627
Ентропія H2 (біграми без перетину): 8.26566
Надлишковість R2 (біграми без перетину): 0.31625

```

За допомогою програми CoolPinkProgram оцінимо значення  $H^{(10)}$ ,  $H^{(20)}$ ,  $H^{(30)}$

$H^{(10)}$

Лабораторная работа №1

Произвольная часть текста:  
учить\_ем

Использованные буквы:

Порядок n-граммы:  
5 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:  
2.49602844092468< H < 3.13622304540502

Двоичная таблица угаданных символов:  
000001000000000000000000000000  
100000000000000000000000000000  
000001000000000000000000000000  
000010000000000000000000000000  
001000000000000000000000000000

Вероятности:  
q[1]=0.38  
q[2]=0.14  
q[3]=0.04  
q[4]=0.06  
q[5]=0.06  
q[6]=0.06  
q[7]=0  
q[8]=0.02  
q[9]=0  
q[10]=0.02  
q[11]=0  
q[12]=0.02  
q[13]=0  
q[14]=0  
q[15]=0  
q[16]=0.06  
q[17]=0  
q[18]=0  
q[19]=0  
q[20]=0  
q[21]=0.02  
q[22]=0  
q[23]=0.04  
q[24]=0.02  
q[25]=0  
q[26]=0  
q[27]=0  
q[28]=0  
q[29]=0  
q[30]=0  
q[31]=0.04  
q[32]=0.02

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

Надлишковість при мінімальному значенні ентропії складає: 0.5008

Надлишковість при максимальному значенні ентропії складає: 0.3728

$H^{(20)}$

Лабораторная работа №1

Произвольная часть текста:  
что\_случилось\_нечто\_непредвиденное\_освобождающее\_его\_от\_необходимости\_выпо

Использованные буквы:

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

Введенный символ: o  
Символ по счету: 1  
Номер эксперимента: 50

Поле ввода символов:  
o  
Продолжить Другой

Неравенство для энтропии:  
1.50123513905468< H < 2.25243127287761

Двоичная таблица угаданных символов:  
000000001000000000000000000000  
000100000000000000000000000000  
000100000000000000000000000000  
000000000000000000000000000010  
00000000000000000000100000000000

Вероятности:  
q[1]=0.6  
q[2]=0.1  
q[3]=0.04  
q[4]=0.06  
q[5]=0.02  
q[6]=0.02  
q[7]=0  
q[8]=0  
q[9]=0.04  
q[10]=0.04  
q[11]=0  
q[12]=0  
q[13]=0  
q[14]=0  
q[15]=0  
q[16]=0  
q[17]=0  
q[18]=0  
q[19]=0.02  
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

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

Надлишковість при мінімальному значенні ентропії складає: 0.6998

Надлишковість при максимальному значенні ентропії складає: 0.5495

$H^{(30)}$

Лабораторная работа №1

Произвольная часть текста:  
то\_его\_поразит\_факт\_насколько\_эти\_учения\_были\_похожи\_друг\_на\_друга\_и\_на\_наш

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  
Поле ввода символов:  

Продолжить Другой

Неравенство для энтропии:  
1,23182379027345 < H < 2,01366068968819  
Двоичная таблица угаданных символов:  
10000000000000000000000000000000  
00000000000000000000000000000000  
01000000000000000000000000000000  
10000000000000000000000000000000  
10000000000000000000000000000000

Вероятности:  
q[1] = 0,64  
q[2] = 0,12  
q[3] = 0,04  
q[4] = 0,04  
q[5] = 0  
q[6] = 0  
q[7] = 0,02  
q[8] = 0,02  
q[9] = 0,04  
q[10] = 0,02  
q[11] = 0  
q[12] = 0,02  
q[13] = 0,02  
q[14] = 0  
q[15] = 0,02  
q[16] = 0  
q[17] = 0  
q[18] = 0  
q[19] = 0  
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  
q[29] = 0  
q[30] = 0  
q[31] = 0  
q[32] = 0

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

Надлишковість при мінімальному значенні ентропії складає: 0.7536

Надлишковість при максимальному значенні ентропії складає: 0.5973

**Висновок:** під час виконання даної лабораторної роботи ми навчилися екпериментально визначати частоти літер і біграм у тексті і на основі цих значень обчислювати ентропію і надлишковість. За допомогою програми CoolPinkProgram ми здійснили серію експериментів, щоб оцінити значення ентропії  $H^{(10)}$ ,  $H^{(20)}$ ,  $H^{(30)}$ .