

**МІНІСТЕРСТВО ОСВІТИ І НАУКИ, МОЛОДІ ТА СПОРТУ 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 гр. ФБ-33

Ольшевський Б.

Степура Н.

Перевірила

Селюх П. В.

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.

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.

Обраний текст - Лев Толстий «Війна і Мир» Том 1 (1.23Мб)

Букви:

З пробілами

=== Літери з пробілами ===		
Символ	Кількість	% від загального
	113782	17.438%
о	61239	9.385%
а	45188	6.925%
е	42479	6.510%
и	35801	5.487%
н	35106	5.380%
т	30595	4.689%
с	28107	4.308%
л	27267	4.179%
в	24799	3.801%
р	24551	3.763%
к	19322	2.961%
д	16381	2.511%
м	15922	2.440%
у	15443	2.367%
п	13839	2.121%
я	12471	1.911%
г	11170	1.712%
ь	10491	1.608%
ы	10224	1.567%
з	9594	1.470%
б	9304	1.426%
ч	7342	1.125%
й	6205	0.951%
ж	5457	0.836%
ш	5087	0.780%
х	4595	0.704%
ю	3494	0.535%
ц	2179	0.334%
э	1629	0.250%
щ	1511	0.232%
ф	1206	0.185%
ё	431	0.066%
ъ	283	0.043%

Без пробілів

Буква, частота

=== Літери без пробілів ===		
Символ	Кількість	% від загального
о	61239	11.368%
а	45188	8.388%
е	42479	7.885%
и	35801	6.646%
н	35106	6.517%
т	30595	5.679%
с	28107	5.217%
л	27267	5.062%
в	24799	4.603%
р	24551	4.557%
к	19322	3.587%
д	16381	3.041%
м	15922	2.956%
у	15443	2.867%
п	13839	2.569%
я	12471	2.315%
г	11170	2.073%
ь	10491	1.947%
ы	10224	1.898%
з	9594	1.781%
б	9304	1.727%
ч	7342	1.363%
й	6205	1.152%
ж	5457	1.013%
ш	5087	0.944%
х	4595	0.853%
ю	3494	0.649%
ц	2179	0.404%
э	1629	0.302%
щ	1511	0.280%
ф	1206	0.224%
ё	431	0.080%
ъ	283	0.053%

Біграми, що перетинаються:

З пробілами

```

=== Біграми з пробілами ===
Символ Кількість % від загального
о 13307 2.039%
и 11383 1.745%
а 10591 1.623%
 10425 1.598%
е 10028 1.537%
с 9856 1.511%
п 9761 1.496%
в 9602 1.472%
н 9346 1.432%
то 8486 1.301%
о 7673 1.176%
я 7092 1.087%
к 7046 1.080%
и 6823 1.046%
сТ 6661 1.021%
ь 6600 1.012%
на 6546 1.003%
го 5753 0.882%
ал 5671 0.869%
не 5541 0.849%
по 5451 0.835%
но 5336 0.818%
ра 5328 0.817%
ко 5257 0.806%
ов 5129 0.786%
л 5031 0.771%
й 4884 0.749%
ка 4809 0.737%
б 4698 0.720%
м 4669 0.716%
во 4471 0.685%
д 4392 0.673%
ро 4381 0.671%
т 4331 0.664%
ер 4330 0.664%
ол 4313 0.661%

```

Повний файл *analysis_results.txt*

Без пробілів

```

=== біграми без пробілів ===
Символ Кількість % від загального
то 8658 1.607%
ст 6836 1.269%
на 6585 1.222%
ов 6482 1.203%
ал 5855 1.087%
го 5799 1.076%
он 5619 1.043%
не 5579 1.036%
но 5541 1.029%
ко 5490 1.019%
по 5454 1.012%
ра 5353 0.994%
ос 5337 0.991%
ен 5072 0.942%
от 4864 0.903%
ка 4861 0.902%
во 4721 0.876%
ло 4599 0.854%
ер 4572 0.849%
ол 4510 0.837%
ро 4435 0.823%
ни 4408 0.818%
ли 4287 0.796%
ан 4252 0.789%
ор 4245 0.788%
ла 4175 0.775%
ел 4012 0.745%
ре 3812 0.708%
пр 3796 0.705%
ва 3764 0.699%
ом 3535 0.656%
ак 3510 0.652%
та 3497 0.649%
за 3481 0.646%
ат 3450 0.640%

```

Повний файл *analysis_results.txt*

```

Ентропія літер з пробілами: 4.3652
Ентропія літер без пробілів: 4.4785
Ентропія біграм з пробілами: 3.9787
Ентропія біграм без пробілів: 4.1569
Ентропія біграм з пробілами: 3.9786
Ентропія біграм без пробілів: 4.1561

Надлишковість літер з пробілами: 0.1346
Надлишковість літер без пробілів: 0.1122
Надлишковість біграм з пробілами: 0.2113
Надлишковість біграм без пробілів: 0.1759
Надлишковість біграм з пробілами: 0.2113
Надлишковість біграм без пробілів: 0.1761

```

2. За допомогою програми CoolPinkProgram оцінити значення H10, H20, H30.

Произвольная часть текста:
и_всегда_
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: 50
Поле ввода символов:
Продолжить Другой

Неравенство для энтропии:
1,63473262855352 < H < 2,35682542858739
Двоичная таблица угаданных символов:
10000000000000000000000000000000
00001000000000000000000000000000
10000000000000000000000000000000
00001000000000000000000000000000
10000000000000000000000000000000

Вероятности:
q[1] = 0,5918367
q[2] = 0,0612244
q[3] = 0,0408163
q[4] = 0,0204081
q[5] = 0,0816326
q[6] = 0,0408163
q[7] = 0
q[8] = 0,0204081
q[9] = 0
q[10] = 0
q[11] = 0
q[12] = 0
q[13] = 0,020408
q[14] = 0,020408
q[15] = 0,020408
q[16] = 0,040816
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] = 0,020408
q[26] = 0
q[27] = 0
q[28] = 0
q[29] = 0
q[30] = 0,020408
q[31] = 0
q[32] = 0

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

$1,63473262855352 < H(10) < 2,35682542858739$

Произвольная часть текста:
ет_с_другими_тeлами_и_организмами_но_тот_закон_который_присущ_только_челове
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:
0,875432250776276 < H < 1,5443814577244
Двоичная таблица угаданных символов:
00100000000000000000000000000000
10000000000000000000000000000000
01000000000000000000000000000000
10000000000000000000000000000000
10000000000000000000000000000000

Вероятности:
q[1] = 0,68
q[2] = 0,16
q[3] = 0,08
q[4] = 0,02
q[5] = 0,02
q[6] = 0,02
q[7] = 0
q[8] = 0
q[9] = 0
q[10] = 0
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
q[20] = 0
q[21] = 0
q[22] = 0
q[23] = 0
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
q[32] = 0

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

$0,875432250776276 < H(20) < 1,54438145772445$

Произвольная часть текста:
ак_же_как_их_не_могут_нарушит

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: 52

Неравенство для энтропии:
 $1,60199746514508 < H < 2,406148748468$

Двоичная таблица угаданных символов:

10000000000000000000000000000000	^
01000000000000000000000000000000	
10000000000000000000000000000000	
00010000000000000000000000000000	
00100000000000000000000000000000	
10000000000000000000000000000000	v

Вероятности:

$q[1] = 0,549019607843137$
 $q[2] = 0,0900392156862745$
 $q[3] = 0,0900392156862745$
 $q[4] = 0,0568235294117647$
 $q[5] = 0,0196078431372549$
 $q[6] = 0,0196078431372549$
 $q[7] = 0,0196078431372549$
 $q[8] = 0$
 $q[9] = 0,0392156862745098$
 $q[10] = 0$
 $q[11] = 0,019607843137254$
 $q[12] = 0$
 $q[13] = 0$
 $q[14] = 0$
 $q[15] = 0,019607843137254$
 $q[16] = 0$
 $q[17] = 0,0392156862745098$
 $q[18] = 0$
 $q[19] = 0$
 $q[20] = 0$
 $q[21] = 0$
 $q[22] = 0$
 $q[23] = 0$
 $q[24] = 0$
 $q[25] = 0$
 $q[26] = 0$
 $q[27] = 0$
 $q[28] = 0$
 $q[29] = 0$
 $q[30] = 0$
 $q[31] = 0,019607843137254$
 $q[32] = 0$

Строка состояния:

$$1,60199746514508 < H(30) < 2,406148748468$$

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

H_0 -ентропія відкритих текстів. Оскільки відкритий текст в нас рівно вірогідний звідси: $H_0 = \log_2 n$, де $n = 32$ літери

$$R = 1 - \frac{H_\infty}{H_0} = 1 - \frac{H_\infty}{\log_2 32} = \frac{H_\infty}{5}$$

$$H^{10}: R_l = 1 - \frac{1,63473262855352}{5} = 0,673053474289296 \approx 0,67$$

$$R_r = 1 - \frac{2,35682542858739}{5} = 0,528634914282522 \approx 0,53$$

$$0,67 < H^{10} < 0,53$$

$$H^{20}: R_l = 1 - \frac{0,875432250776276}{5} = 0,8249135498447448 \approx 0,82$$

$$R_r = 1 - \frac{1,54438145772445}{5} = 0,69112370845511 \approx 0,69$$

$$0,82 < H^{20} < 0,69$$

$$H^{30}: R_l = 1 - \frac{1,60199746514508}{5} = 0,679600506970984 \approx 0,67$$

$$R_r = 1 - \frac{2,406148748468}{5} = 0,51877024503064 \approx 0,52$$

$$0.67 < H^{30} < 0.52$$

Висновки: під час роботи над даним комп'ютерним практикумом ми мали змогу дослідити поняття ентропії, біграм та надлишковості російської мови за допомогою власноруч написаного скрипту та програми CoolPinkProgram, що дало змогу оцінити різницю вихідних даних на практиці у порівнянні з отриманими внаслідок експериментального «вгадування» наступних літер(людського фактору у тому числі)