

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

**КОМП'ЮТЕРНИЙ ПРАКТИКУМ №1**  
**з дисципліни «Криптографія»**

Виконав:

студент 3 курсу

НН ФТІ групи ФБ-25

Черняк Денис

**Тема:** «Експериментальна оцінка ентропії на символ джерела відкритого тексту»

**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.

### Хід роботи

#### Завдання 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.

### Виконання

```
Ентропія для тексту з пробілами (включно з пробілами в H1):  
H1 (ентропія букв): 4.406696524751244  
H2 (ентропія біграм): 4.014339899570864  
  
Ентропія для тексту без пробілів:  
H1 (ентропія букв): 4.48359558232663  
H2 (ентропія біграм): 4.178746528108722
```

Lab1.py  bigram_matrix_with_spaces.csv							
	Перша літера	а	б	в	г	д	
1	а	1.8013274397594536e-05	0.0010184428217101527	0.004116726018096413	0.0016960190663581317	0.0029098366334575788	
2	б	0.0008355388047499619	0.00012470728429103908	6.789618811401017e-05	1.3856364921226565e-05	1.8013274397594536e-05	
3	в	0.008179412213000042	1.5242001413349222e-05	8.729509900372736e-05	0.00016073383308622816	0.00047665895329019383	
4	г	0.006297717856697474	2.771272984245313e-06	0.00021615929277113443	0.00041984785711316494	0.001504801230445205	
5	д	0.006861671908991395	6.789618811401017e-05	0.0011583921074145409	4.295473125580235e-05	0.00017597583449957739	
6	е	0.0001053083734013219	0.0023306405797503082	0.0022585874821599303	0.005064501378708309	0.003610968698471643	
7	ж	0.001823497623633416	1.662763790547188e-05	4.15690947636797e-06	3.0484002826698444e-05	0.0009380759051670385	
8	з	0.008471781512837923	0.00017597583449957739	0.0010641688259502003	0.0007025177015061868	0.0011916473832254847	
9	и	0.00026049966051905945	0.0006983607920298189	0.003094126286909892	0.0007413155232856212	0.0025994540592221035	
10	й	1.3856364921226565e-06	1.3856364921226565e-06	1.5242001413349222e-05	4.9882913716415634e-05	0.00028128420790089927	
11	к	0.009513780154914159	2.771272984245313e-06	0.0002549571145505688	0.00010807964638556722	1.3856364921226565e-06	
12	л	0.009820005819673266	7.898128005099142e-05	4.849727722429298e-05	0.00018290401696019067	0.00090759190234034	
13	м	0.004097327107206695	0.0008895786279427455	4.4340367747925006e-05	0.0001718189250232094	0.00010946528287768987	
14	н	0.01599855893804819	4.15690947636797e-06	1.2470728429103908e-05	0.0002327869306766063	0.0012678573902922307	
15	о	8.31381895273594e-06	0.00515318211420416	0.008202968033366127	0.00614945475204035	0.006002577283875348	
16	п	0.0031537086560711664	0.0	0.0	9.699455444858596e-06	0.0	
17	р	0.010320220593329547	6.651055162188752e-05	0.0005944380551206196	0.00016766201554684144	0.0006609486067425072	
18	с	0.001937119815987474	0.00010807964638556722	0.0017057185218029903	0.00013302110324377503	0.00035056603250703213	
19	т	0.007270434674167579	5.265418670066095e-05	0.003213291025324403	6.235364214551954e-05	0.0001662763790547188	
20	у	3.8797821779434386e-05	0.0006886613365849603	0.0013080408485637877	0.0013967215840596377	0.002254430572683562	
21	ф	0.00016211946957835082	0.0	0.0	0.0	0.0	
22	х	0.0006581773337582618	2.771272984245313e-06	0.000557025869833308	1.9398910889717193e-05	6.928182460613283e-06	
23	ц	0.0006166082389945822	0.0	7.066746109825548e-05	2.771272984245313e-06	4.15690947636797e-06	
24	ч	0.002359738946084884	0.0	1.3856364921226565e-06	4.15690947636797e-06	0.0	

Lab1.pybigram\_matrix\_with\_spaces.csvbigram\_matrix\_without\_spaces.csv

	Перша літера	а	б	в	г	д
1	а	0.00023565174055307126	0.0014781790998329015	0.004935156308137765	0.0027150688576640936	0.0032269630692004303
2	б	0.000688914897023572	0.00010598690723439568	7.103377825283966e-05	1.4657763766458978e-05	2.142288550482466e-05
3	в	0.006715510845617667	0.0002153563753379742	0.0004250751492273103	0.0005637601448638068	0.000644941605724195
4	г	0.005136982439999008	4.735585216855977e-05	0.00026271222750653397	0.0003720816956101125	0.0012819905694202967
5	д	0.00560377583994624	0.00010147682607548523	0.0010598690723439567	0.00013417491447758603	0.00020408117244069806
6	е	0.00023339669997361602	0.002459121751895925	0.0032799565228176283	0.00471077977048197	0.0036261052517640056
7	ж	0.0014950919041788156	3.2698088402100796e-05	3.382560869182841e-05	3.2698088402100796e-05	0.0007813715607812362
8	з	0.0069128268963199995	0.0001916784492536943	0.0010260434636521283	0.0006505792071728331	0.0010452113085774979
9	и	0.0004656658796575044	0.0012707153665230207	0.0048348070023520075	0.0017093207592270624	0.003108573438779031
10	й	9.13291434679367e-05	0.00022775909852497796	0.0006968075390516652	0.0004239476289375827	0.0005795454289199934
11	к	0.00778214503969999	0.00025481958547844067	0.0005964582332659076	0.0003033029579367281	0.00021986645649688466
12	л	0.008097850720823722	0.00030894055938536616	0.000927949198445826	0.001269587846233293	0.0011489431752324383
13	м	0.003426534160482218	0.0009538821651095612	0.0007667137970147772	0.0005389546984897993	0.0004983639680596053
14	н	0.013063450076784132	0.00023677926084279886	0.0005085116506671537	0.0003292359246004632	0.0012267420752236436
15	о	0.00013304739418785842	0.0050918816284099035	0.008785638097557566	0.005816877174704759	0.005843937661658222
16	п	0.0025662361794200486	1.1275202897276136e-06	1.1275202897276136e-06	1.0147682607548524e-05	3.382560869182841e-06
17	р	0.008422576564265274	0.00010034930578575762	0.0006212636796399151	0.0001804032463564182	0.0006111159970323666
18	с	0.0016315218592358569	0.00022663157823525035	0.0016957905157503309	0.0003980146622738476	0.0005051290897979709
19	т	0.005988260258743356	0.00028300759272163105	0.003143526567760587	0.00022888661881470556	0.00036757161445120207
20	у	0.00013868499563649648	0.0007667137970147772	0.001721723482414066	0.0015244074317117336	0.0021783691997537494
21	ф	0.00013417491447758603	1.1275202897276136e-06	9.020162317820909e-06	1.1275202897276136e-06	0.0
22	х	0.0005739078274713553	0.00010260434636521285	0.0007926467636785124	0.00016687300287968683	0.00017589316519750774
23	ц	0.0005073841303774261	7.892642028093296e-06	7.554385941175012e-05	2.142288550482466e-05	2.142288550482466e-05
24	ч	0.001925804654854764	4.510081158910454e-06	2.0295365215097047e-05	4.9610892748015e-05	5.6376014486380684e-06

Lab1.pyletter\_frequencies\_with\_spaces.csv

	Літера	Частота
1		0.1570284617175879
2	о	0.09288623391929589
3	а	0.07072610883791221
4	е	0.06176130936256968
5	и	0.05827880032125766
6	н	0.05418894322388712
7	т	0.048463333380856656
8	л	0.045465562224661756
9	р	0.045352456718134426
10	с	0.043892540264134625
11	в	0.035166307865585035
12	м	0.028336256017640657
13	к	0.027700393968339964
14	д	0.025730077035305075
15	у	0.02461422943309429
16	п	0.023390028656563207
17	г	0.020132780161864434
18	я	0.01671205143924381
19	ь	0.01553822538410725
20	з	0.015374745156185398
21	ы	0.01479401016048626
22	б	0.013583115914134861
23	ч	0.011703093293033557
24	й	0.008225336583928563
25	ш	0.0081846665269681

Lab1.py letter\_frequencies\_with\_spaces.csv letter\_frequencies\_without\_spaces.csv

	Літера	Частота
1	о	0.11018905111381966
2	а	0.08390094519919315
3	е	0.07326618581738928
4	и	0.06913495613387259
5	н	0.06428324179758102
6	т	0.05749106723057651
7	л	0.05393487224645762
8	р	0.0538006974832648
9	с	0.05206882827096086
10	в	0.041717076162782175
11	м	0.03361472449636544
12	к	0.03286041427303775
13	д	0.030523067347838488
14	у	0.02919936002020514
15	п	0.027747115524471108
16	г	0.023883107848321632
17	я	0.0198251669010027
18	ь	0.018432680913245306
19	з	0.018238747642075854
20	ы	0.01754983352181693
21	б	0.016113374292340877
22	ч	0.013883141673892185
23	й	0.009757549585467633
24	ж	0.00970906626767527
25	ш	0.008835239028394311

## Завдання 2

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

Під час виконання були проблеми з ієрогліфами, і на жаль, повністю прибрати її не вдалось, але залишок символів не заважав виконання роботи, оскільки одна кнопка відповідала за продовження дії, а інша до перехід до наступного значення

### Виконання

$$H^{(10)}$$

$$2,1984629707794 < H < 2,92598127152669$$

Лабораторная работа №1

Произвольная часть текста:  
едствления\_о\_том\_что\_мы\_называем\_добром\_то\_хотя\_нам\_и\_пришлось\_бы\_воевать\_

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: 1

Номер эксперимента: 53

Неравенство для энтропии:  
 $2,1984629707794 < H < 2,92598127152669$

Двоичная таблица угаданных символов:  
00001000000000000000000000000000  
00000100000000000000000000000000  
00000000000010000000000000000000  
10000000000000000000000000000000  
00000000000000000100000000000000

Вероятности:  
q[1] = 0,4716981  
q[2] = 0,1320754  
q[3] = 0  
q[4] = 0,0566037  
q[5] = 0,0377358  
q[6] = 0,0377358  
q[7] = 0  
q[8] = 0  
q[9] = 0  
q[10] = 0  
q[11] = 0,018867  
q[12] = 0  
q[13] = 0,037735  
q[14] = 0,018867  
q[15] = 0,018867  
q[16] = 0,018867  
q[17] = 0,018867  
q[18] = 0,037735  
q[19] = 0  
q[20] = 0  
q[21] = 0,018867  
q[22] = 0  
q[23] = 0,018867  
q[24] = 0  
q[25] = 0,018867  
q[26] = 0,018867  
q[27] = 0  
q[28] = 0  
q[29] = 0  
q[30] = 0,018867  
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



$$2,62138549320994 < H < 3,24023073433794$$

Лабораторная работа №1

---

Произвольная часть текста:  
ричины\_он\_делает\_ви

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

Поле ввода символов:

Продолжить Другой

Неравенство для энтропии:  
2,62138549320994 < H < 3,24023073433794

Двоичная таблица угаданных символов:  
10000000000000000000000000000000  
10000000000000000000000000000000 |  
10000000000000000000000000000000  
10000000000000000000000000000000  
10000000000000000000000000000000  
10000000000000000000000000000000  
10000000000000000000000000000000  
~~~~~

Вероятности:  
q[1] = 0,4509803  
q[2] = 0,0392156  
q[3] = 0,0196078  
q[4] = 0,0392156  
q[5] = 0  
q[6] = 0  
q[7] = 0,0196078  
q[8] = 0  
q[9] = 0,0392156  
q[10] = 0,039215  
q[11] = 0,039215  
q[12] = 0,019607  
q[13] = 0,019607  
q[14] = 0,039215  
q[15] = 0,039215  
q[16] = 0,019607  
**q[17] = 0,019607**  
q[18] = 0,039215  
q[19] = 0  
q[20] = 0,039215  
q[21] = 0,019607  
q[22] = 0  
q[23] = 0,039215  
q[24] = 0,019607  
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:

$$2,33839612073776 < H < 2,92891056032524$$

### Завдання 3

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

#### Виконання

Для обрахунків використаємо цю формулу

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

$H_{(0)}$  можна знайти за наступною формулою, 32 це кількість літер у нашому алфавіті, це вказано у методичці, що у наданому тексті використовується лише 32 символи

$$H_{(0)} = \log_2 32 = 5$$

$$H^{(10)}$$

$$R = 1 - \frac{2,1984629707794}{5} \approx 0,56030740584412$$

$$R = 1 - \frac{2,92598127152669}{5} \approx 0,414803745694662$$

$$H^{(20)}$$

$$R = 1 - \frac{2,62138549320994}{5} \approx 0,475722901358012$$

$$R = 1 - \frac{3,24023073433794}{5} \approx 0,351953853132412$$

$$H^{(30)}$$

$$R = 1 - \frac{2,33839612073776}{5} \approx 0,532320775852448$$

$$R = 1 - \frac{2,92891056032524}{5} \approx 0,414217887934952$$

## Висновки

У ході виконання роботи я дослідив та оцінив ентропію для символів джерела відкритого тексту російською мовою. Було розроблено програму для підрахунку частот символів та біграм у тексті, а також обчислено ентропії  $H_1$  та  $H_2$  як з урахуванням пробілів, так і без них. Це дало можливість більш детально вивчити вплив пробілів на загальну ентропію тексту. Зокрема, результати показали, що включення пробілів впливає на розподіл частот окремих символів та біграм, проте не призводить до суттєвих змін значень  $H_1$  та  $H_2$ , що свідчить про низький внесок пробілів у загальну ентропію при обробці тексту значного обсягу.

Додатково, за допомогою програми CoolPinkProgram було обчислено ентропії  $H_{10}$ ,  $H_{20}$ ,  $H_{30}$  для тексту, що включає лише літери та пробіли, без урахування розділових знаків. Це дозволило оцінити надлишковість російської мови, яка, за нашими підрахунками, варіюється від 47% до 65% залежно від значення ентропії  $H_{NN}$ . Високий рівень надлишковості свідчить про значну передбачуваність структури російської мови, що можна враховувати при проектуванні систем стиснення тексту або криптографічного аналізу.

Результати також вказали на помітну різницю у значеннях  $H$  між текстом із пробілами та без них, а також між різними завданнями, що підтверджує залежність ентропії від обраного тексту та характеру символів, які він містить.