

МІНІСТЕРСТВО ОСВІТИ І НАУКИ 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:  
ФБ-33 Охріменко Анастасія

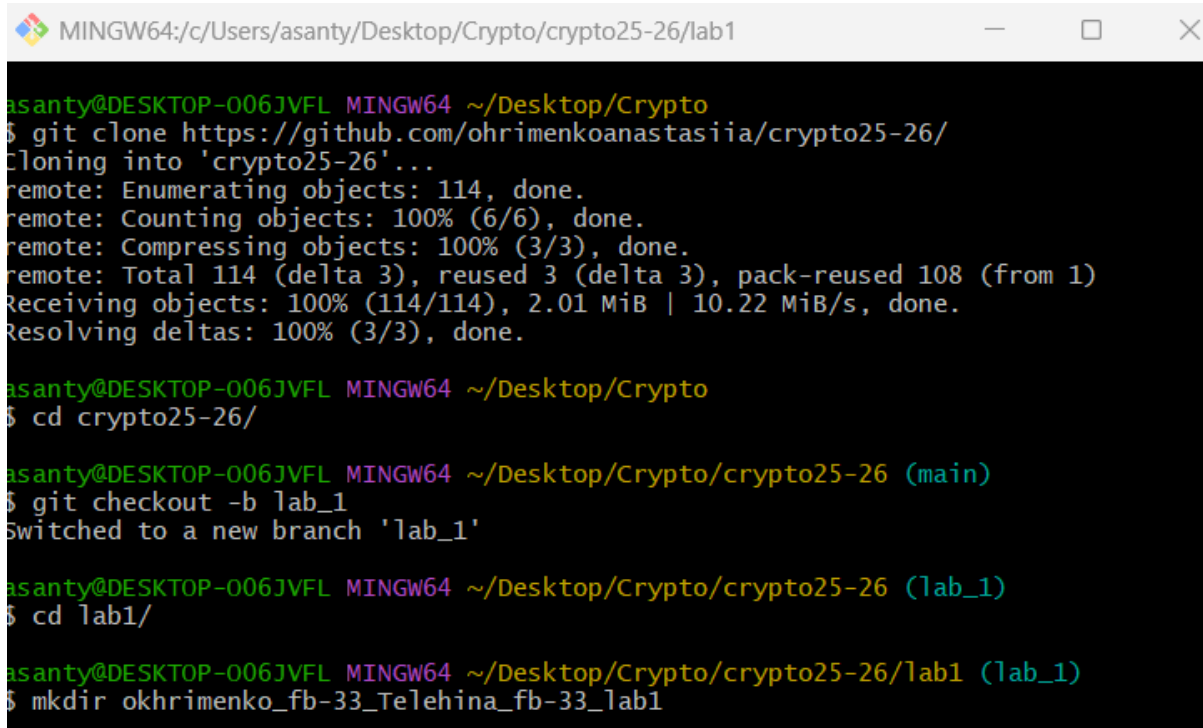
ФБ-33 Телегіна Софія

Перевірила:  
Селюх Поліна Валентинівна

**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

Для початку налаштувала гіт:



```
MINGW64:/c/Users/asanty/Desktop/Crypto/crypto25-26/lab1
asanty@DESKTOP-006JVFL MINGW64 ~/Desktop/Crypto
$ git clone https://github.com/ohrimenkoanastasiia/crypto25-26/
Cloning into 'crypto25-26'...
remote: Enumerating objects: 114, done.
remote: Counting objects: 100% (6/6), done.
remote: Compressing objects: 100% (3/3), done.
remote: Total 114 (delta 3), reused 3 (delta 3), pack-reused 108 (from 1)
Receiving objects: 100% (114/114), 2.01 MiB | 10.22 MiB/s, done.
Resolving deltas: 100% (3/3), done.

asanty@DESKTOP-006JVFL MINGW64 ~/Desktop/Crypto
$ cd crypto25-26/

asanty@DESKTOP-006JVFL MINGW64 ~/Desktop/Crypto/crypto25-26 (main)
$ git checkout -b lab_1
Switched to a new branch 'lab_1'

asanty@DESKTOP-006JVFL MINGW64 ~/Desktop/Crypto/crypto25-26 (lab_1)
$ cd lab1/

asanty@DESKTOP-006JVFL MINGW64 ~/Desktop/Crypto/crypto25-26/lab1 (lab_1)
$ mkdir okhrimenko_fb-33_Telehina_fb-33_lab1
```

**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.**

Написала спочатку просто підрахунок  $H_1$ , перевірила на простому прикладі, все ок, закомітила.

Далі дописала вивід кожного символу та частоту його появи.

Додала вираховування частоти біграм та знаходження  $H_2$ .

Аби вийшов файл з текстом на  $> 1$ Мб скопіювали текст книжок про Гулівера, Тома Сойера, Війни світів. Побачила далі що потрібно було робити разом з пробілами біграми також, до цього весь код був без них, тому код умовно поділений на дві частини: розрахунок без пробілів і з пробілами. Ще одне завдання порахувати біграми з кроком два. Переробила код для виводу результату у файл csv.

Результат:

WITHOUT SPACES: Symbol Frequency Table				
Symbol	Count	Probability		
e	127190	0.125003		
t	96527	0.094867		
a	81245	0.079848		
o	77309	0.07598		
n	70260	0.069052		
i	68039	0.066869		
h	63806	0.062709		
s	62155	0.061086		
r	58509	0.057503		

WITHOUT SPACES: Bigram Matrix (Overlapping)													
	a	b	c	d	e	f	g	h	i	j	k	l	
a	0.00005	0.002201	0.003429	0.004649	0.000011	0.001113	0.002005	0.000522	0.00313	0.000208	0.000989	0.005	
b	0.000904	0.000165	0.000001	0.000024	0.004737	0.000007	0.000002	0.000036	0.000541	0.000096	0	0.002	
c	0.002704	0.000012	0.000463	0.00001	0.003973	0.00001	0.000012	0.004181	0.001014	0.000002	0.001997	0.	
d	0.004694	0.001962	0.000816	0.001158	0.005791	0.00117	0.000848	0.002058	0.005296	0.000137	0.000108	0.00	
e	0.011373	0.002286	0.0045	0.01252	0.004704	0.002922	0.001713	0.002814	0.004681	0.000165	0.000429	0.005	
f	0.002342	0.000309	0.000265	0.000182	0.001968	0.001142	0.000168	0.000659	0.002179	0.00004	0.000018	0.000	
g	0.002452	0.000305	0.000195	0.000216	0.002893	0.000288	0.0004	0.003232	0.001455	0.00002	0.000009	0.000	
h	0.010424	0.000253	0.000176	0.000157	0.02895	0.0002	0.000117	0.000514	0.008925	0.00002	0.000018	0.000	
i	0.001123	0.000653	0.003505	0.003087	0.00216	0.001527	0.002422	0.000669	0.00008	0.000018	0.000535	0.003	
j	0.000087	0	0	0	0.000393	0	0	0	0.000032	0	0		
k	0.00045	0.000097	0.000048	0.00005	0.002377	0.000128	0.000017	0.000158	0.001323	0.000005	0.000007	0.000	
l	0.004344	0.000283	0.000333	0.003027	0.006969	0.000899	0.000133	0.000298	0.004529	0.000017	0.000351	0.005	
m	0.004568	0.000771	0.00012	0.00013	0.007057	0.000197	0.000078	0.000291	0.002497	0.000021	0.000021	0.00	
n	0.003729	0.00063	0.002742	0.014462	0.00597	0.000857	0.009301	0.00131	0.003396	0.000257	0.00058	0.000	
o	0.001456	0.001326	0.001034	0.001871	0.000579	0.008473	0.000725	0.000957	0.001124	0.000071	0.001187	0.002	
p	0.002139	0.000058	0.000018	0.000016	0.003478	0.000055	0.000008	0.000028	0.001329	0.00001	0.000006	0.001	
q	0	0	0	0	0	0	0	0	0	0	0	0	

WITHOUT SPACES: Bigram Matrix (Non-overlapping)													
	a	b	c	d	e	f	g	h	i	j	k	l	
a	0.000049	0.002231	0.003412	0.004621	0.00001	0.001093	0.001956	0.000533	0.003084	0.000193	0.000953	0.005897	
b	0.000924	0.000171	0.000002	0.000024	0.004806	0.000008	0.000002	0.000026	0.000503	0.000086	0	0.002011	
c	0.002713	0.000012	0.000448	0.000008	0.003955	0.000012	0.000014	0.004145	0.001054	0	0.002036	0.000977	
d	0.004757	0.001956	0.000871	0.001154	0.005761	0.00116	0.000843	0.002082	0.00536	0.000143	0.000114	0.001059	
e	0.011412	0.002201	0.004486	0.012513	0.004678	0.002921	0.001728	0.002764	0.004694	0.000173	0.00046	0.005107	
f	0.002361	0.000299	0.000228	0.000187	0.00193	0.001148	0.000177	0.000649	0.002117	0.000033	0.00002	0.000871	
g	0.002498	0.000291	0.000199	0.000224	0.002939	0.000271	0.000411	0.003151	0.001457	0.000026	0.000008	0.000788	
h	0.010302	0.000263	0.000195	0.000151	0.028908	0.000193	0.000106	0.000541	0.008731	0.000022	0.000018	0.000163	
i	0.001136	0.00066	0.003573	0.003003	0.00219	0.001569	0.002428	0.000686	0.000075	0.000014	0.000564	0.00323	
j	0.000081	0	0	0	0.000427	0	0	0	0.000028	0	0	0	
k	0.000409	0.000088	0.000053	0.000039	0.002394	0.000136	0.000016	0.000155	0.001254	0.000004	0.000006	0.000149	
l	0.004393	0.000271	0.000297	0.002988	0.006933	0.000877	0.00014	0.000358	0.004529	0.000016	0.000358	0.005272	
m	0.004666	0.000802	0.000104	0.000126	0.007084	0.000189	0.000051	0.000293	0.002488	0.000024	0.000018	0.000098	
n	0.003778	0.000641	0.002701	0.014286	0.005956	0.000861	0.009195	0.001311	0.003401	0.000256	0.000607	0.000757	
o	0.001482	0.001321	0.001032	0.001877	0.000554	0.008564	0.000704	0.001006	0.001156	0.000063	0.001179	0.002292	
p	0.002099	0.000053	0.000016	0.000018	0.003566	0.000057	0.000008	0.000267	0.001236	0.000012	0.000004	0.001742	
q	0	0	0	0	0	0	0	0	0	0	0	0	
r	0.005152	0.000739	0.001126	0.002217	0.014255	0.000772	0.000655	0.000977	0.00491	0.000045	0.00058	0.000863	
s	0.006276	0.001087	0.001887	0.000574	0.00813	0.001152	0.000419	0.004413	0.005118	0.000112	0.000427	0.001105	

WITH SPACES: Symbol Frequency Table		
Symbol	Count	Probability
_	243437	0.193061
e	127190	0.10087
t	96527	0.076552
a	81245	0.064433
o	77309	0.061311
n	70260	0.055721
i	68039	0.053959
h	63806	0.050602
s	62155	0.049293
r	58509	0.046401
d	47405	0.037595
l	38898	0.030849
u	29157	0.023123
m	27282	0.021636
c	24792	0.019662
w	24498	0.019429

WITH SPACES: Bigram Matrix (Overlapping)												
	_	a	b	c	d	e	f	g	h	i	j	l
_	0.007444	0.022675	0.008704	0.00718	0.005371	0.003578	0.006894	0.003507	0.011922	0.012952	0.000701	
a	0.004881	0.000005	0.001488	0.002337	0.003512	0.000002	0.000584	0.00133	0.000161	0.002506	0.000144	
b	0.000085	0.000707	0.00013	0	0.000018	0.003822	0.000002	0.000001	0.000021	0.000433	0.000078	
c	0.00022	0.002149	0	0.000364	0.000003	0.003198	0	0	0.003362	0.000807	0	
d	0.023972	0.000896	0.00004	0.000013	0.000336	0.004271	0.000035	0.000247	0.000021	0.002453	0.00001	
e	0.035638	0.005482	0.000151	0.001822	0.008912	0.003044	0.000893	0.000528	0.000254	0.001659	0.000023	
f	0.007324	0.001105	0.00001	0.000002	0	0.001437	0.000758	0.000001	0.000003	0.001366	0.000001	
g	0.006038	0.001104	0.000013	0.000005	0.000057	0.002218	0.000002	0.000254	0.002288	0.000726	0.000001	
h	0.005222	0.007494	0.00005	0.000007	0.000013	0.02328	0.000016	0.000002	0.000007	0.006667	0.000001	
i	0.003965	0.000692	0.000427	0.00242	0.002209	0.001692	0.001088	0.001871	0.000007	0.000044	0	
j	0	0.000071	0	0	0	0.000317	0	0	0	0.000026	0	
k	0.001939	0.000029	0.000006	0.000006	0.000007	0.001894	0.000032	0	0.000006	0.000891	0.000001	
l	0.004233	0.002989	0.000036	0.000067	0.002321	0.005538	0.000581	0.000033	0.000021	0.003309	0	
m	0.003576	0.003175	0.000479	0.000004	0.000041	0.005658	0.000049	0.000004	0.000007	0.001719	0	
n	0.013588	0.000934	0.000034	0.001883	0.011415	0.004425	0.000295	0.007379	0.000155	0.001698	0.000124	
o	0.008291	0.000579	0.0005	0.000477	0.001132	0.000269	0.006552	0.000309	0.000116	0.000561	0.000033	
p	0.001213	0.001481	0.000005	0.000001	0	0.002799	0.000007	0	0.000165	0.000932	0.000002	
q	0	0	0	0	0	0	0	0	0	0	0	
r	0.010769	0.002669	0.000151	0.000471	0.001565	0.01133	0.000182	0.000385	0.000111	0.003254	0.000002	
s	0.019899	0.001844	0.000033	0.00092	0.000033	0.006029	0.00007	0.000028	0.002493	0.002412	0.000002	
t	0.019624	0.002229	0.000015	0.000391	0.000007	0.005693	0.000048	0.000002	0.025988	0.004754	0.000002	

WITH SPACES: Bigram Matrix (Non-overlapping)											
	_	a	b	c	d	e	f	g	h	i	j
_	0.00733	0.022856	0.008911	0.007157	0.005329	0.003608	0.007087	0.003443	0.011972	0.012987	0.000671
a	0.004785	0.000002	0.001464	0.002357	0.003604	0.000003	0.000568	0.00127	0.000154	0.00249	0.000148
b	0.000098	0.000726	0.000143	0	0.000019	0.003762	0.000002	0	0.000024	0.000431	0.000073
c	0.000224	0.002165	0	0.000328	0	0.003233	0	0	0.003367	0.000828	0
d	0.023933	0.000906	0.000048	0.000014	0.000317	0.004257	0.000027	0.000254	0.000013	0.002466	0.000011
e	0.035463	0.005502	0.000144	0.001757	0.008879	0.003072	0.000914	0.000508	0.000257	0.001662	0.000005
f	0.007385	0.001071	0.000008	0.000003	0	0.001391	0.000719	0	0.000003	0.00135	0
g	0.006043	0.001169	0.000013	0.000003	0.000048	0.002275	0.000005	0.000252	0.002292	0.000693	0.000002
h	0.005229	0.00748	0.000046	0.000008	0.000008	0.023503	0.000013	0	0.000008	0.006632	0.000002
i	0.003972	0.000671	0.000419	0.002362	0.002165	0.001711	0.00109	0.001886	0.000002	0.00004	0
j	0	0.000089	0	0	0	0.00032	0	0	0	0.000027	0
k	0.001884	0.000033	0.000008	0.000008	0.000005	0.001854	0.000038	0	0.000008	0.000871	0.000002
l	0.004192	0.002944	0.000046	0.000067	0.0023	0.005336	0.000574	0.00003	0.000014	0.003298	0
m	0.003488	0.003193	0.000444	0.000008	0.000036	0.005704	0.000049	0.000005	0.000008	0.001711	0
n	0.013514	0.000934	0.000038	0.001826	0.011545	0.00436	0.000287	0.00739	0.000149	0.001683	0.000127
o	0.008284	0.000581	0.000496	0.000508	0.001115	0.000274	0.006557	0.000309	0.000111	0.000558	0.000038
p	0.001247	0.001505	0.000005	0.000002	0	0.002811	0.000013	0	0.00016	0.000901	0.000002
q	0	0	0	0	0	0	0	0	0	0	0

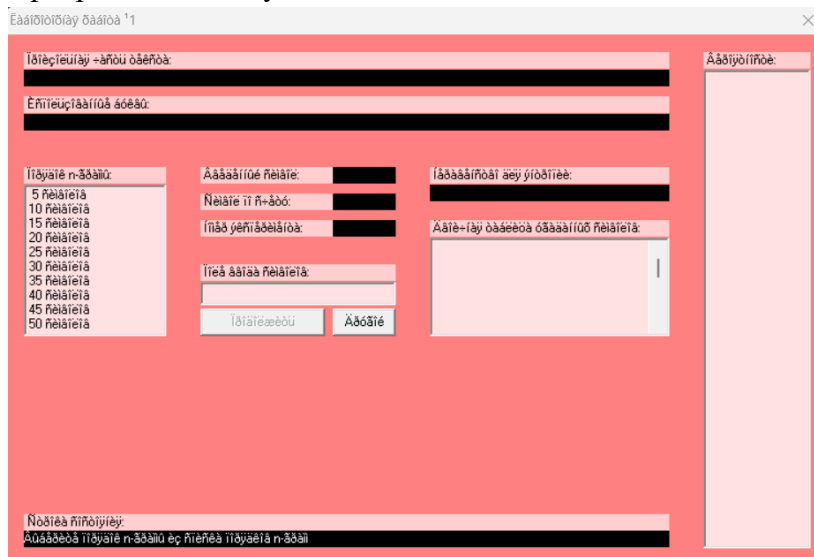
WITHOUT SPACES SUMMARY		
H1	4.173785	
H2 overlap	3.882399	
H2 non-ov	3.881669	

WITH SPACES SUMMARY		
H1	4.07582	
H2 overlap	3.703747	
H2 non-ov	3.702527	

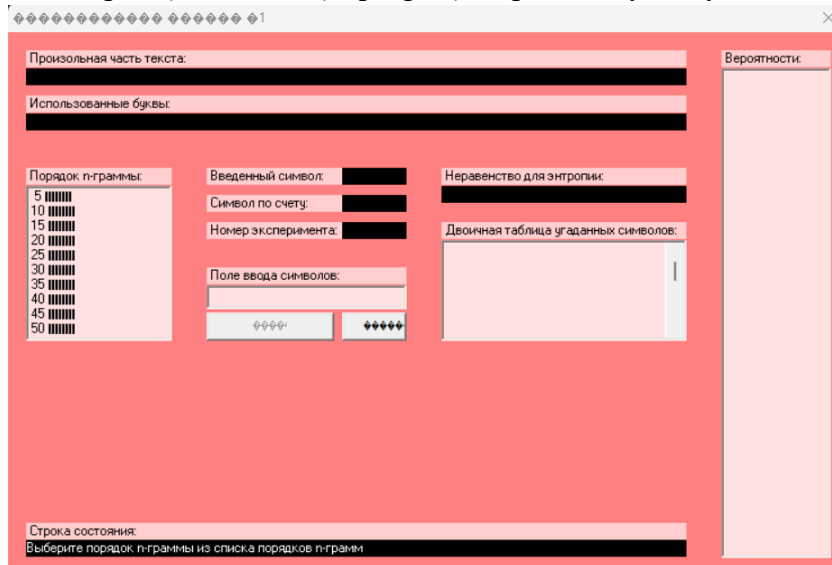
Бачимо що значення з пробілами та без не сильно відрізняються. Бачимо що  $H2 < H1$ , тобто є закономірності в тексті. З пробілами ентропія знижується, тобто кожна літера/біграма несе в собі більше інформації. Також дійсно різниця між  $H2$  з перекриттям та без не сильно відрізняється

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

Програма не показувала звичайні символи:



Переключила тимчасово системну локаль (визначає, яке кодування використовується для старих (не-Unicode) програм) на російську мову, виглядає краще:



Для H(10):

H20:

H30:

Произвольная часть текста:  
даю\_в\_моей\_голове\_сразу\_же\_во

Использованные буквы:

Порядок n-грамм:  
5 10 15 20 25 35 40 45 50

Введенный символ:

Символ по счету:

Номер эксперимента: 51

Поле ввода символов:

Неравенство для энтропии:  
 $1.75564662760354 < H < 2.5386885704511$

Двоичная таблица угаданных символов:

Вероятности:

q[1]	= 0.48
q[2]	= 0.2
q[3]	= 0.04
q[4]	= 0.02
q[5]	= 0.04
q[6]	= 0
q[7]	= 0.08
q[8]	= 0.02
q[9]	= 0.02
q[10]	= 0.02
q[11]	= 0
q[12]	= 0
q[13]	= 0.02
q[14]	= 0
q[15]	= 0.02
q[16]	= 0
q[17]	= 0.02
q[18]	= 0
q[19]	= 0
q[20]	= 0
q[21]	= 0.02
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:

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

Відкритий текст буде рівно вірогідний, тому  $H_0 = \log_2 n = \log_2 32 = 5$

$$R = 1 - \frac{H}{H_0}$$

H10:

$$R_1 = 1 - \frac{1,667}{5} = 0,666$$

$$R_2 = 1 - \frac{2,412}{5} = 0,5176$$

$$0.666 < H(10) < 0.5176$$

H20:

$$R_1 = 1 - \frac{1,3773}{5} = 0,72454$$

$$R_2 = 1 - \frac{2,1756}{5} = 0,56488$$

$$0.72454 < H(20) < 0.56488$$

H30:

$$R_1 = 1 - \frac{1,7556}{5} = 0,64888$$

$$R_2 = 1 - \frac{1}{5} = 0,49226$$

$$0.64888 < H(30) < 0.49226$$



**Висновки:** набули практичних навичок при дослідженні ентропій та порівняння їх між собою. Дійсно певні літери зустрічаються частіше (наприклад а, е), також є закономірності і в біграмах (наприклад, після пробілу з меншою вірогідністю буде йти “й” ніж с або в). За допомогою CoolPinkProgram можна було легко дослідити які символи зустрічаються частіше за інші. Також варто зазначити що при використанні CoolPinkProgram певна частина результатів залежила від попередньої букви, тобто якщо попереднім був пробіл то існує багато варіацій наступної літери.