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

КРИПТОГРАФІЯ КОМП’ЮТЕРНИЙ ПРАКТИКУМ No1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконали: Бойко Т. Я.

Хандрос А. В.

Група: ФБ-02

Київ – 2022

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.

Завдання

Створити програму для розрахування частот букв та біграм у тексті достатньої довжини російською мовою, а також ентропії за її безпосереднім означенням. Застосувати цю програму обрахувавши частоти букв та біграм, ентропію на тексті. Оцінити надлишковість мови на основі отриманих результатів.

Хід роботи

- Частота букв

	freq with spaces	freq without spaces
	0,1628	
о	0,089	0,1062
а	0,0801	0,0957
е	0,0699	0,0835
н	0,0545	0,0652
и	0,0529	0,0632
т	0,0496	0,0592
л	0,046	0,055
с	0,0437	0,0522
р	0,0372	0,0444
в	0,0331	0,0395
к	0,0296	0,0353
у	0,028	0,0335
м	0,0264	0,0316
п	0,0234	0,0279
д	0,0227	0,0271
я	0,02	0,0239
ь	0,0186	0,0222
ы	0,0154	0,0184
з	0,0152	0,0182
г	0,0126	0,0151
б	0,0126	0,015
ч	0,0119	0,0142
й	0,0087	0,0104
ж	0,0082	0,0098
ш	0,0068	0,0082
х	0,0061	0,0073
ю	0,0044	0,0052
щ	0,0031	0,0038
ц	0,0029	0,0035
ф	0,0025	0,003
э	0,0019	0,0023
ъ	0,0002	0,0002

- Частота біграм (повний список в .xlsx файлі)

	перехр. з пробілом		перехр. без пробіла
а	0,026124567	ла	0,017653808
о	0,019833486	то	0,01494615
е	0,018268913	на	0,013452062
п	0,016211607	но	0,011604409
н	0,015639748	по	0,011541151
и	0,01522835	ст	0,011407481
ла	0,014670046	ен	0,011000071
я	0,013480932	ал	0,01056894
с	0,013423872	не	0,01042661
в	0,013007115	ко	0,009609907
ь	0,012385447	ос	0,009388128
то	0,012120008	ра	0,009244292
о	0,011551302	ас	0,009167856
на	0,011167015	ро	0,0089457
по	0,009655088	ов	0,008879054
но	0,009569025	он	0,008667442
ст	0,009322817	от	0,008613597
не	0,00866521	ни	0,008476539
т	0,008502858	ол	0,007687323
и	0,008450211	ат	0,007500562
ал	0,008161129	ка	0,007495291

	з пробілом		без пробіла
а	0,013061338	ла	0,008785673
о	0,009918004	то	0,007488889
е	0,009088903	на	0,006760296
п	0,008033139	по	0,005757585
н	0,007796703	но	0,005738005
и	0,007603772	ст	0,005673994
ла	0,007339909	ен	0,005483092
я	0,006711306	ал	0,005292565
с	0,006673161	не	0,005190524
в	0,006524994	ко	0,004830181
ь	0,00617507	ос	0,004708184
то	0,006050862	ас	0,004621958
о	0,005775336	ра	0,004620828
на	0,005559706	ро	0,004503726
по	0,004866162	ов	0,0044544
но	0,004754564	он	0,004305669
ст	0,00467323	от	0,004295879
не	0,004363342	ни	0,00421643
т	0,004246385	ол	0,003830483
и	0,004229992	ка	0,003758565
ал	0,004070162	ат	0,003755929

- Ентропія та надлишковість

Ентропія H1 без пробілів 4.4583236806704765

Надлишковість R1 без пробілів 0.1000914373693188

Ентропія H1 з пробілами 4.4583236806704765

Надлишковість R1 з пробілами 0.10833526386590475

Ентропія H1 без пробілів 4.369566780679863

Ентропія H2 без пробілів, біграми пересікаються 4.646183743845846

Надлишковість R2 без пробілів, біграми пересікаються 0.062172055212114996

Ентропія H2 з пробілами, біграми пересікаються 4.478967238346061

Надлишковість R2 з пробілами, біграми пересікаються 0.1042065523307878

Ентропія H2 без пробілів, біграми не пересікаються 4.646370191547851

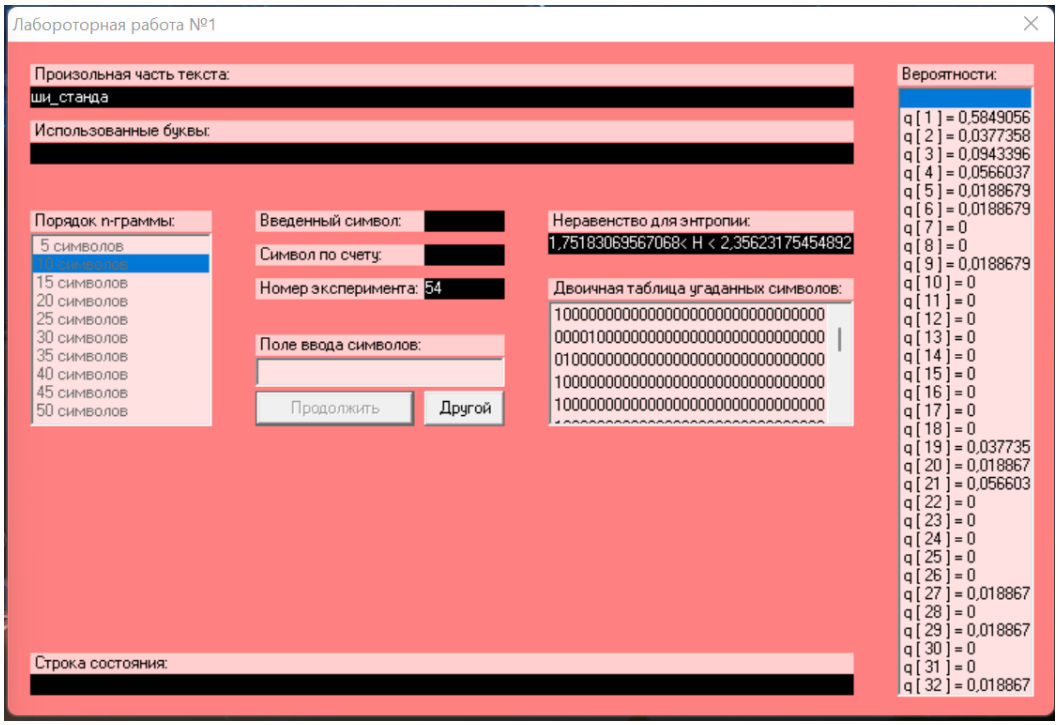
Надлишковість R2 без пробілів, біграми не пересікаються 0.06213442091376997

Ентропія H2 з пробілами, біграми пересікаються 4.47947639932785

Надлишковість R2 з пробілами, біграми пересікаються 0.10410472013443006

- CoolPinkProgram

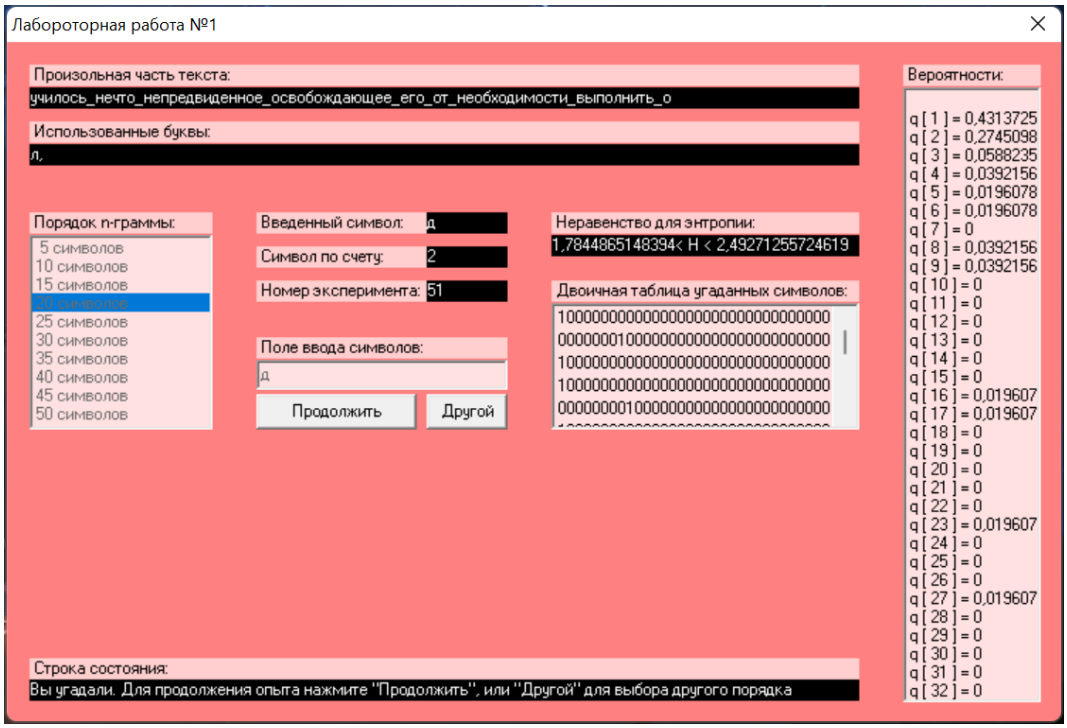
H(10)



1.752<H<2.356

0.6496 > R > 0.5288

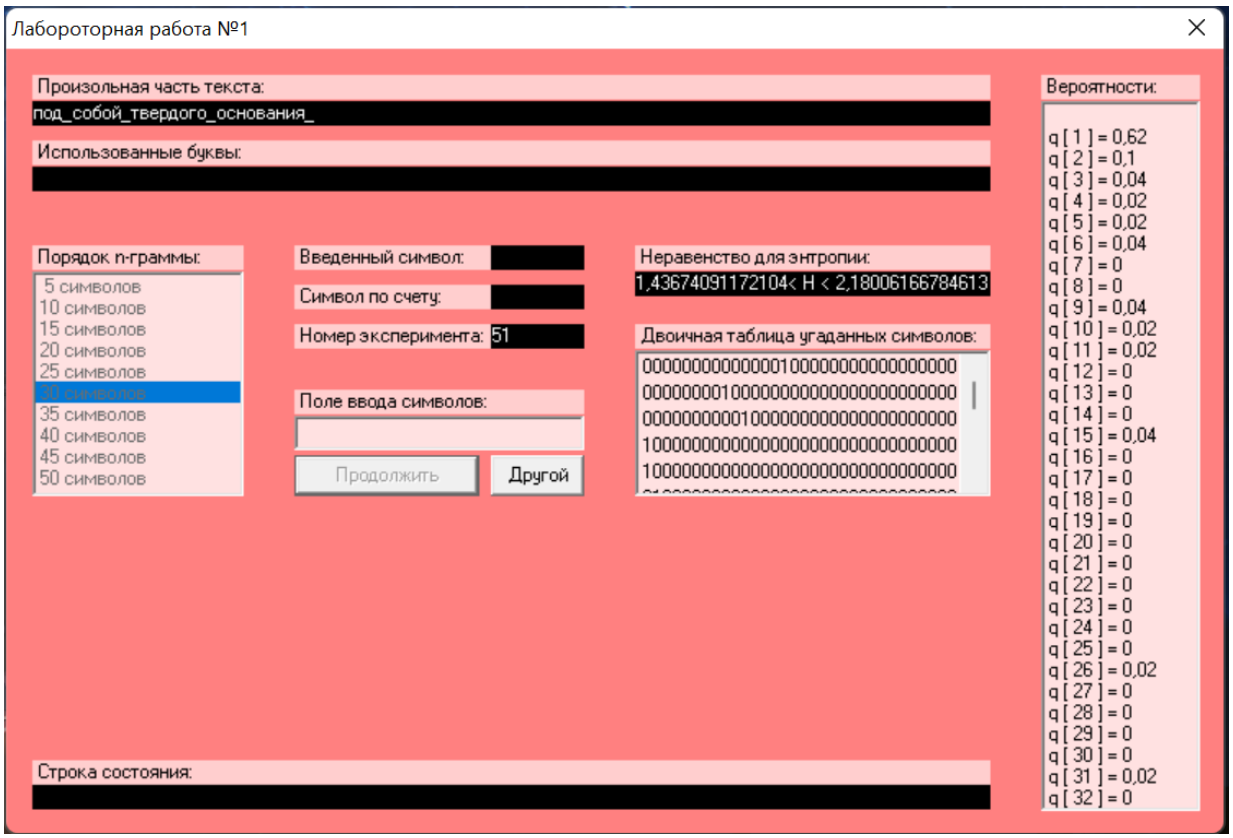
H(20)



1.784<H<2.492

0.6432 > R > 0.5016

H(30)



1.437<H<2.18

0.7126 > R > 0.564

- Висновок

Виконуючи лабораторну роботу ми дослідили ентропію та надлишковість російської мови на різних джерелах тексту. В теорії, результати не повинні сильно відрізнятися для різних джерел тексту, але на даному етапі код містить помилку, через яку перехресні біграми у тексті з пробілами обробляються некоректно (помилково обчислюється ентропія, а отже і надлишковість). Відловити помилку не зламавши всього іншого не вдалось.