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

Фізико-технічний інститут

КОМП’ЮТЕРНИЙ ПРАКТИКУМ №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконав: Маслюк В.О. ФБ-25

Київ 2023

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.

Хід роботи

1. **Підрахунок частоти, ентропії та надлишковості для букв**

Текст з пробілами

Всього унікальних літер: 32

буква	частота
	0.1296
о	0.0926
е	0.0782
и	0.0720
а	0.0662
н	0.0646
т	0.0600
с	0.0539
р	0.0440
в	0.0364
л	0.0352
м	0.0269
д	0.0262
к	0.0252
п	0.0195
у	0.0180
ы	0.0169
ь	0.0158
я	0.0155
з	0.0150
г	0.0127
ч	0.0116
й	0.0111
б	0.0099
х	0.0097
ю	0.0065
ц	0.0063
ж	0.0062
ш	0.0049
э	0.0040
щ	0.0030
ф	0.0025

Ентропія літер в тексті з пробілами: 4.4054

Надлишковість: 0.1189

Текст без пробілів
Всього унікальних літер: 31

буква	частота
о	0.1064
е	0.0898
и	0.0827
а	0.0760
н	0.0742
т	0.0690
с	0.0619
р	0.0506
в	0.0419
л	0.0404
м	0.0309
д	0.0301
к	0.0289
п	0.0224
у	0.0206
ы	0.0194
ь	0.0182
я	0.0178
з	0.0172
г	0.0146
ч	0.0133
й	0.0127
б	0.0113
х	0.0111
ю	0.0075
ц	0.0073
ж	0.0071
ш	0.0057
э	0.0047
щ	0.0034
ф	0.0028

Ентропія літер в тексті без пробілів: 4.4221
Надлишковість: 0.1156

2. Підрахунок частоти, ентропії та надлишковості для біграм

Текст з пробілами (біграми, що перетинаються):

Всього унікальних біграм (з пробілами, перекриваючі): 476

біграма	частота
а	0,0141
г	0,0016
го	0,0051
од	0,0074
ды	0,0009
ы	0,0046
ш	0,0005
шл	0,0007
ли	0,0060
и	0,0169
д	0,0076
да	0,0046
б	0,0035
бы	0,0012
ыс	0,0011
ст	0,0180
тр	0,0076
ро	0,0090
о	0,0143
и	0,0088
н	0,0127
не	0,0093
ес	0,0065
сл	0,0040
лы	0,0004
ыш	0,0004
шн	0,0005
но	0,0151
к	0,0049
ка	0,0048
ак	0,0032
к	0,0018
п	0,0107
по	0,0079

Повністю в xlsx файлі

Ентропія біграм: 3.9493

Надлишковість біграм: 0.2101

Текст без пробілів (біграми, що перетинаються):
Всього унікальних біграм (без пробілів, перекриваючі): 506

біграма	частота
аг	0,0006
го	0,0059
од	0,0093
ды	0,0010
ыш	0,0006
шл	0,0008
ли	0,0073
ид	0,0024
да	0,0053
аш	0,0006
иб	0,0006

Ентропія біграм: 4.0208
Надлишковість біграм: 0.1958

Текст з пробілами (біграми, що не перетинаються):
Всього унікальних біграм (з пробілами, без перекриття): 404

біграма	частота
а	0,0162
го	0,0046
ды	0,0004
ш	0,0007
ли	0,0067
д	0,0088
шл	0,0011
и	0,0201
бы	0,0011
ст	0,0190
ро	0,0099
и	0,0085
н	0,0127
ес	0,0060
лы	0,0004
шн	0,0011

Ентропія біграм: 3.9117

Надлишковість біграм: 0.2177

Текст без пробілів (біграми, що не перетинаються):

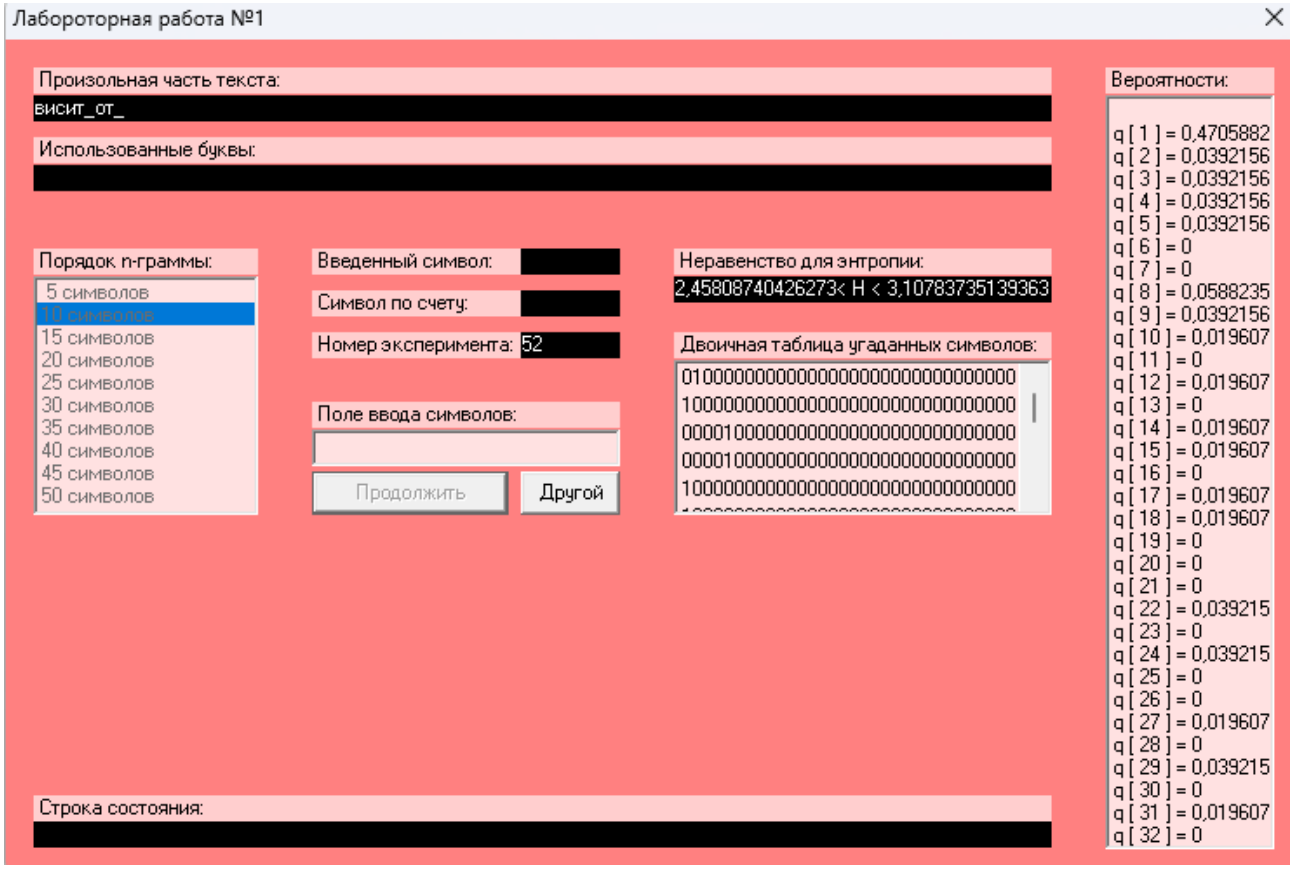
Всього унікальних біграм (без пробілів, без перекриття): 420

біграма	частота
аг	0,0004
од	0,0089
ыш	0,0012
ли	0,0081
да	0,0057
шл	0,0004
иб	0,0008
ыс	0,0016
тр	0,0113
ои	0,0024
не	0,0113
сл	0,0044
но	0,0186
ка	0,0049

Ентропія біграм: 3.9749
Надлишковість біграм: 0.2050

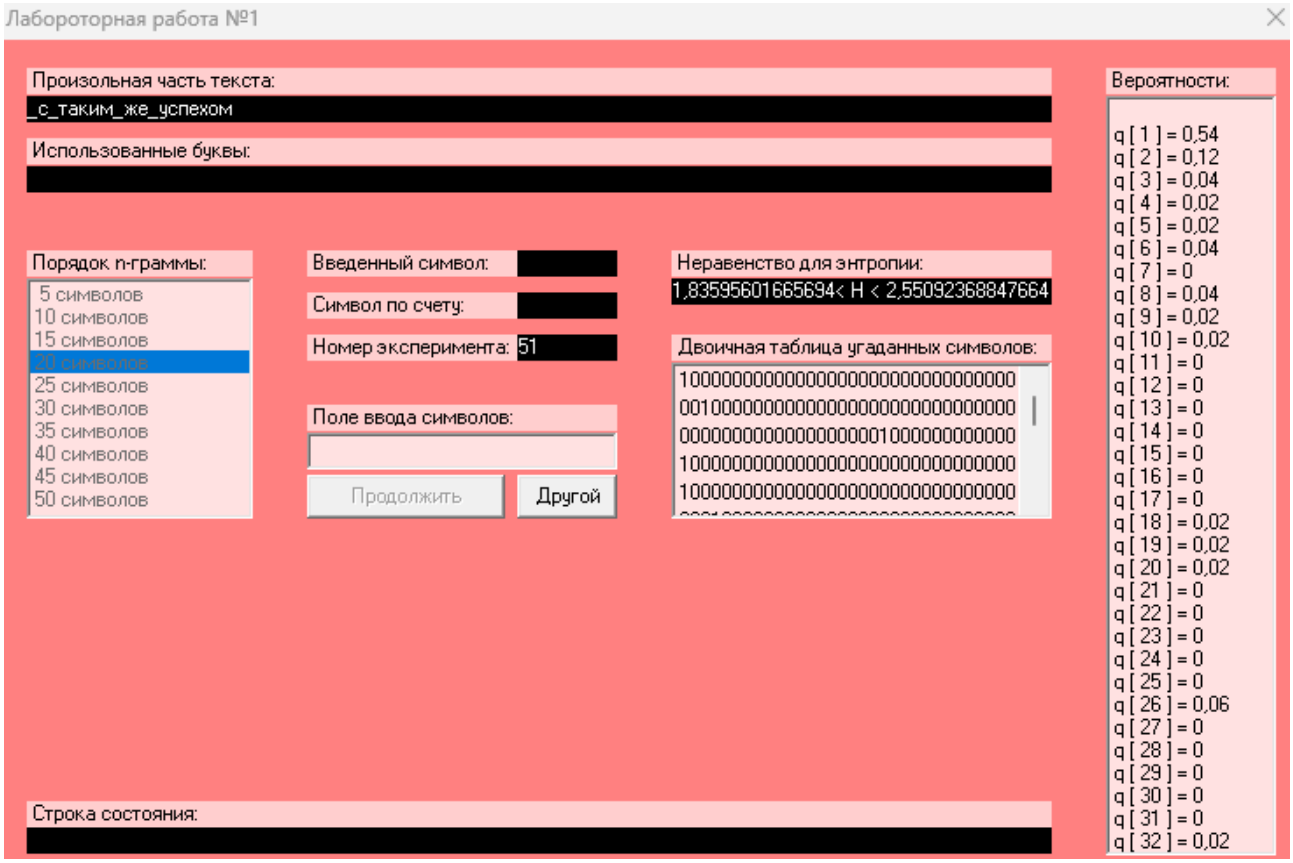
3. Оцінка значень $H^{(10)}$, $H^{(20)}$, $H^{(30)}$ з використанням програми CoolPinkProgram

Результати експерименту для $H^{(10)}$



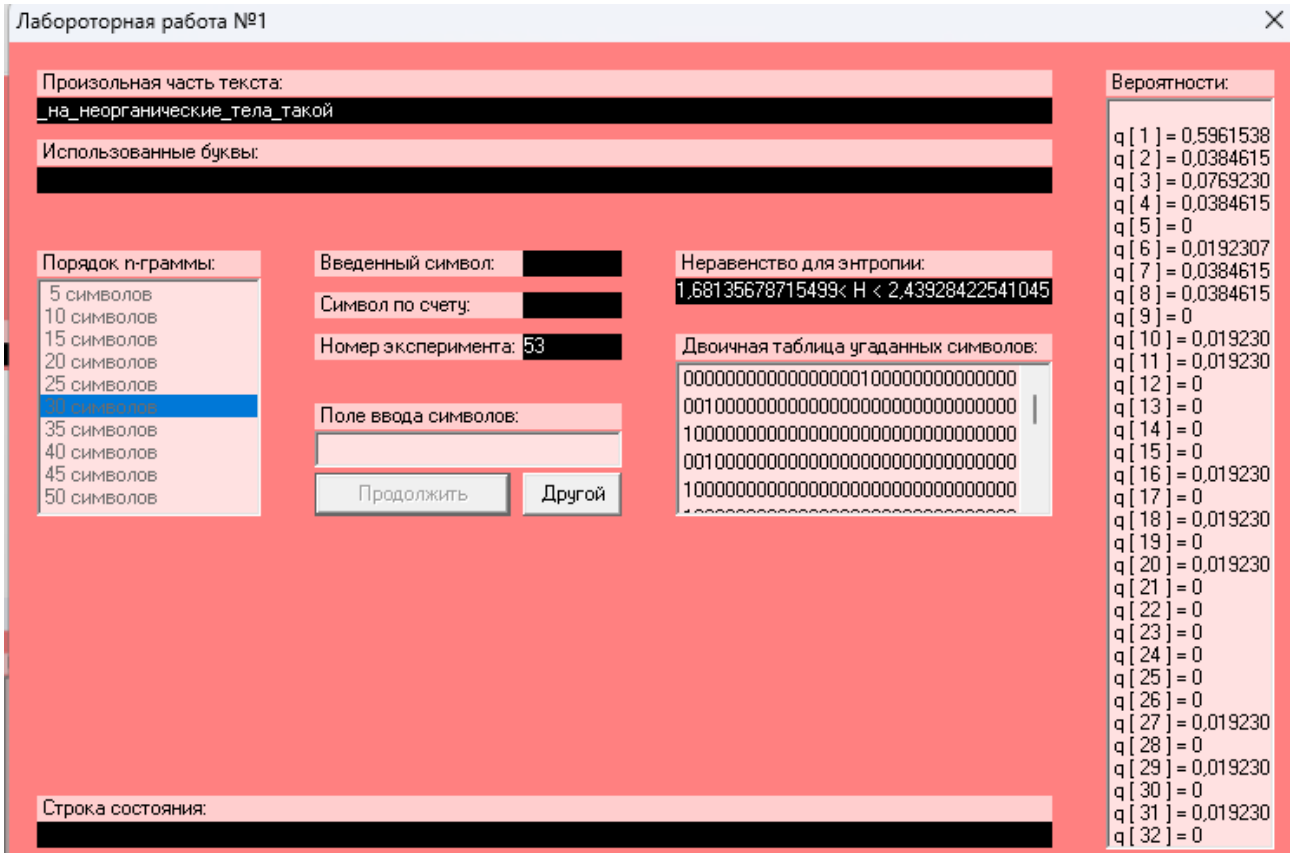
Надлишковість: 0,44340

Результати експерименту для $H^{(20)}$



Надлишковість: 0,5613

Результати експерименту для $H^{(30)}$



Надлишковість: 0,5679

Висновок: Під час виконання лабораторної роботи ми навчилися екпериментально визначати частоти літер і біграм у тексті і на основі цих значень обчислювати ентропію і надлишковіть у різних моделях відкритого тексту. Також, за допомогою спеціальної програми приблизно обчислили значення H(10), H(20), H(30)