КРИПТОГРАФІЯ
КОМП’ЮТЕРНИЙ ПРАКТИКУМ №1
Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконав студент: Медвецький Давид
Група: ФБ-13

Мета роботи
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
1. Написати програми для підрахунку частот букв і частот біграм в тексті, а також підрахунку H1 та H2 за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення H1 та H2 на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення H1 та H2 на тому ж тексті, в якому вилучено всі пробіли.

Частота букв:

ж	0.008557
и	0.057537
з	0.013583
н	0.055102
ь	0.017617
" "	0.160293
е	0.070295
о	0.092549
б	0.015554
ы	0.016313
к	0.027134
в	0.036694
у	0.023118
д	0.027289
т	0.049512
л	0.043768
п	0.023706
р	0.037738
ю	0.004322
ч	0.011326
я	0.018395
а	0.066974
м	0.027475
й	0.009273
ш	0.007013
г	0.016665
ц	0.002900
с	0.045070
э	0.003097
х	0.007694

ф	0.000939
щ	0.002500

Частота біграм топ 20 (перетинаються):

о_	0.019691
и_	0.019047
е_	0.016700
_п	0.016530
_н	0.015449
а_	0.015060
_с	0.014242
_в	0.014192
я_	0.011892
то	0.011405
ст	0.011115
_о	0.011006
ь_	0.010914
на	0.010167
_и	0.009487
не	0.009347
но	0.009187
_м	0.008577
по	0.008542
_к	0.008381

Частота біграм топ 20 (не перетинаються):

о_	0.019442
и_	0.019012
е_	0.016771
_п	0.016482
_н	0.015491
а_	0.014952
_с	0.014368
_в	0.014130
я_	0.011884
то	0.011603
_о	0.011022
ст	0.010986
ь_	0.010846
на	0.010278
_и	0.009502
не	0.009319

но	0.009304
_м	0.008603
по	0.008558
_к	0.008409

Текст з пробілами:

H_1 = 4.381057 $R_1 = 1 - \frac{4.381075}{5} = 0.123785$

H_2 = 3.978438 (перетинаються) $R_2 = 1 - \frac{3.978438}{5} = 0.2043124$

H_2 = 3.978091 (не перетинаються) $R_2 = 1 - \frac{3.978438}{5} = 0.2043818$

Текст без пробілів:

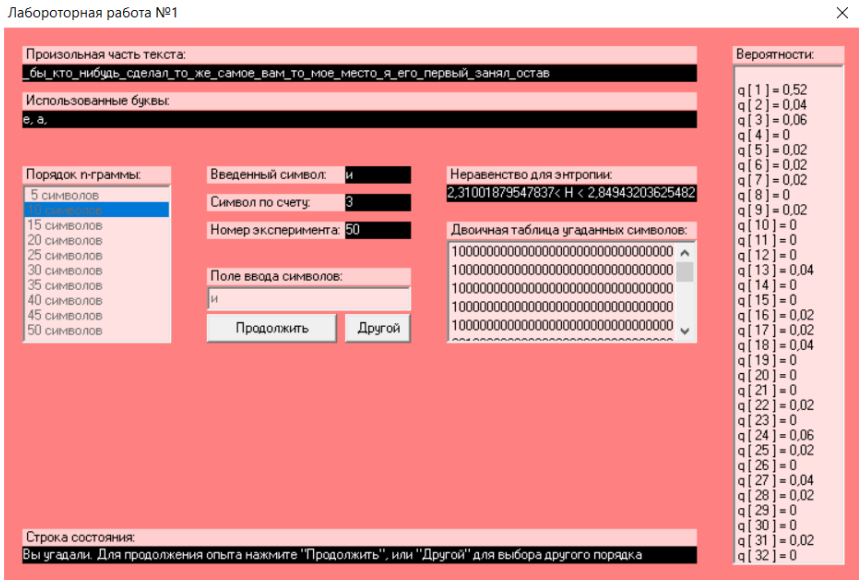
H_1 = 4.461136 $R_1 = 1 - \frac{4.461136}{5} = 0.1077728$

H_2 = 4.147518 (перетинаються) $R_2 = 1 - \frac{4.147518}{5} = 0.1704964$

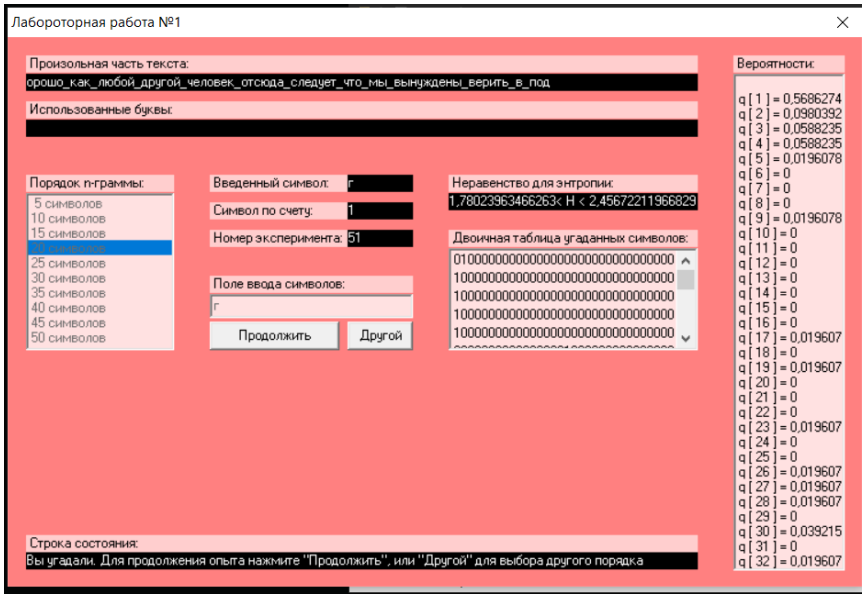
H_2 = 4.147599 (не перетинаються) $R_2 = 1 - \frac{4.147599}{5} = 0.1704802$

2. За допомогою програми CoolPinkProgram оцінити значення

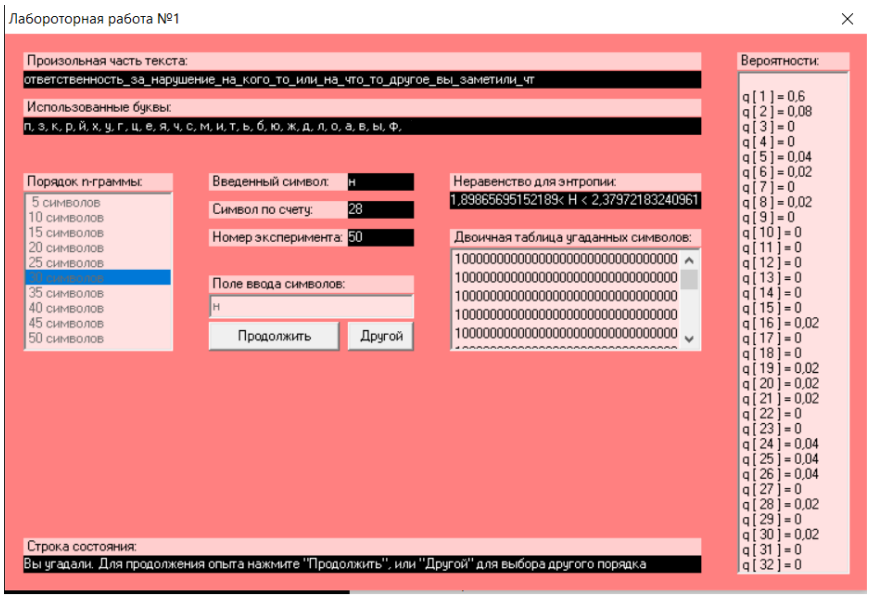
а) 2.31001879547837 < $H^{(10)}$ < 2.84943203625482



б) 1.77023963466263 < $H^{(20)}$ < 2.45672211966929



в) $1.89865695152189 < H^{(30)} < 2.37972183240961$



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

Скористаємося формулою: $R = 1 - \frac{H_\infty}{H_0}$ де $H_0 = \log 2\,(32) = 5$

$1 - \frac{2.31001879547837}{5} < R < 1 - \frac{2.84943203625482}{5}$ $0.537997 < R < 0.430113$

$1 - \frac{1.77023963466263}{5} < R < 1 - \frac{2.45672211966929}{5}$ $0.645952 < R < 0.508655$

$1 - \frac{1.89865695152189}{5} < R < 1 - \frac{2.37972183240961}{5}$ $0.620268 < R < 0.524055$

Деякі труднощі:

1. Було важко знайти нормальний текст, який б відповідав потрібному розміру, тому склепав його самостійно з 3 книжок

2. В деяких книжках, наприклад я хотів використати “Преступление и наказание”, глави нумеруються римськими цифрами, тому цей твір я відкинув. У “Приключения Робинзона Крузо” були власні назви рослин написані латинськими літерами, їх видаляв вручну.

3. Трохи було не зрозуміло в чому суть біграм що перетинаються і що не перетинаються

4. Пропустив ділення на 2 при підрахунку ентропії для біграм

Висновок

У ході виконання лабороторної роботи ми ознайомилися з поняттями єнтропії на символ джерела та його надлишковості. Написали программу яка опрацьовує і аналізує текст. Попрацювали з CoolPinkProgram за допомогою якої оцінили приблизне значення ентропії. Порівняли різні моделі джерела відкритого тексту, щоб вибрати ту, яка найкраще відображає реальну структуру даних.