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

Фізико-технічний інститут

СИМЕТРИЧНА КРИПТОГРАФІЯ

КОМП’ЮТЕРНИЙ ПРАКТИКУМ №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконали:
студентки групи ФІ-94
Зацаренко А. Ю.
Футурська О.В.

Перевірив:
Чорний О.М.

Київ – 2022

ЗМІСТ

ЗАГАЛЬНІ ВІДОМОСТІ

1. Мета комп’ютерного практикуму

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.

2. Постановка задачі

Створити програму для експериментальної оцінки ентропії на символ джерела відкритого тексту, порівняти різні моделі джерела відкритого тексту для наближеного визначення ентропії.

3. Хід роботи

1. Уважно прочитати методичні вказівки до виконання комп’ютерного практикуму.
2. Написати програми для підрахунку частот букв і частот біграм в тексті, а також підрахунку H_1та H_2 за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення H_1та H_2 на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення H_1та H_2 на тому ж тексті, в якому вилучено всі пробіли.
3. За допомогою програми CoolPinkProgram оцінити значення $H^{(10)}$, $H^{(20)}$, $H^{(30)}$.
4. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

4. Опис труднощів

Реалізуючи програмний код, ми зіштовхнулися з проблемою зайвих символів у тексті при зчитувані літер на межі строк. В ASSCI дані символи розташовані на 10 місці (LINE FEED) та на 13 (CARRIAGE RETURN). В результаті ми вирішили вчинити таким чином: видаляти символи, а на їх місцях записувати пробіли. На коректність програми це не впливає, адже потім ми використовуємо функцію Regex.Replace з бібліотеки System.Text.RegularExpressions, яка замінює усі пробіли одним.

Ще одна проблема з реалізацією була в тому, що текст у файлі написаний російською мовою і зчитувався з файлу не правильно. Як виявилося, проблема полягала в кодуванні файлу і достатньо було при зчитувані «перетворити» його, використовуючи функцію Encoding.GetEncoding(866).

ПРАКТИЧНА ЧАСТИНА

1. Програмний код

```
using System;
using System.Collections.Generic;
using System.Linq;
using System.Text;
using System.Threading.Tasks;
using System.Threading;
using System.Text.RegularExpressions;
using System.IO;

namespace Lab1
{
    class Program
    {
        static void Main(string[] args)
        {
            string alph = "абвгдежзийклмнопрстуфхцчшщыьэюя ";
            int n = alph.Length;
            FilterText();
            var sr = new StreamReader("TEXT", Encoding.GetEncoding(866));
            double[] s = Symbol(sr, alph);
            for (int i = 0; i < n; i++)
                Console.WriteLine("{0}  -  {1:N5}", alph[i], s[i]);
            sr = new StreamReader("TEXT", Encoding.GetEncoding(866));
            double[,] b = Biggram(sr, alph);
            for (int i = 0; i < n; i++)
            {
                for (int j = 0; j < n; j++)
                {
                    if (b[i,j] != 0)
                        Console.WriteLine("{0}{1}  -  {2:N5}", alph[i], alph[j], b[i, j]);
                }
            }
            sr = new StreamReader("TEXT", Encoding.GetEncoding(866));
            double[,] b2 = BiggramIS(sr, alph);
            for (int i = 0; i < n; i++)
            {
                for (int j = 0; j < n; j++)
                {
                    if (b2[i, j] != 0)
                        Console.WriteLine("{0}{1}  -  {2:N5}", alph[i], alph[j], b2[i, j]);
                }
            }
            Console.WriteLine("H1 = {0:N5}", Entropy(s, n));
            Console.WriteLine("H2 = {0:N5}", Entropy(b, n));
            Console.WriteLine("H3 = {0:N5}", Entropy(b2, n));
            Console.WriteLine("H1\\\" \" = {0:N5}", Entropy(s, n-1));
            Console.WriteLine("H2\\\" \" = {0:N5}", Entropy(b, n - 1));
            Console.WriteLine("H3\\\" \" = {0:N5}", Entropy(b2, n - 1));
            Console.ReadKey();
            sr.Close();
        }

        static double[,] Biggram(StreamReader sr, string alph)
        {
            int flag, i = 0, j = 0, sum = 0;
            char l1, l2;
            int n = alph.Length;
            double[,] count = new double[n,n];
            Array.Clear(count, 0, n);
            while (!sr.EndOfStream)
            {
```

```
            l1 = (char)sr.Read();
            l2 = (char)sr.Read();
            flag = 0;
            for (int k = 0; k < n; k++)
            {
                if (alph[k] == l1)
                {
                    i = k;
                    if (flag == 1)
                        break;
                    else flag++;
                }
                if (alph[k] == l2)
                {
                    j = k;
                    if (flag == 1)
                        break;
                    else flag++;
                }
            }
            count[i, j]++;
            sum++;
        }
        for (i = 0; i < n; i++)
        {
            for (j = 0; j < n; j++)
                count[i, j] = count[i, j] / sum;
        }
        return count;
    }

    static double[,] BiggramIS(StreamReader sr, string alph)
    {
        int i = 0, j = 0, sum = 0;
        char l1, l2;
        int n = alph.Length;
        double[,] count = new double[n, n];
        Array.Clear(count, 0, n);
        l2 = (char)sr.Read();
        for (j = 0; j < n; j++)
        {
            if (l2 == alph[j])
                break;
        }
        while (!sr.EndOfStream)
        {
            l1 = l2;
            i = j;
            l2 = (char)sr.Read();
            for (j = 0; j < n; j++)
            {
                if (l2 == alph[j])
                    break;
            }
            count[i, j]++;
            sum++;
        }
        for (i = 0; i < n; i++)
        {
            for (j = 0; j < n; j++)
                count[i, j] = count[i, j] / sum;
        }
        return count;
    }

    static double[] Symbol(StreamReader sr, string alph)
```

```
        {
            int k, sum = 0;
            char letter;
            int n = alph.Length;
            double[] count = new double[n];
            Array.Clear(count, 0, n);
            while (!sr.EndOfStream)
            {
                k = sr.Read();
                letter = (char)k;
                for (int i = 0; i < n; i++)
                {
                    if (letter == alph[i])
                    {
                        count[i]++;
                        sum++;
                        break;
                    }
                }
            }
            for (int i = 0; i < n; i++)
                count[i] = count[i] / sum;
            return count;
        }

        static double Entropy(double[,] count, int n)
        {
            double h = 0;
            for (int i = n - 1; i >= 0; i--)
            {
                for (int j = n - 1; j >= 0; j--)
                {
                    if (count[i, j] != 0)
                        h += (count[i, j] * Math.Log(count[i, j], 2));
                }
            }
            return -h;
        }

        static double Entropy(double[] count, int n)
        {
            double h = 0;
            for (int i = n - 1; i >= 0; i--)
                h += (count[i] * Math.Log(count[i], 2));
            return -h;
        }

        static void FilterText()
        {
            var text = File.ReadAllText("TEXT", Encoding.GetEncoding(866));
            int index;
            text = text.ToLower();
            for (int i = 0; i < text.Length; i++)
            {
                index = (int)text[i];
                if ((index < 1072 || index > 1103) && index != 32)
                    text = text.Remove(i, 1).Insert(i, " ");
                if (index == 1098)
                    text = text.Remove(i, 1).Insert(i, "ь");
            }
            text = Regex.Replace(text, @"\s+", " ");
            File.WriteAllText("TEXT", text, Encoding.GetEncoding(866));
        }
    }
}
```

2. Результати підрахунку частот

Символ	Частота	Символ	Частота
а	0,06735	р	0,03322
б	0,01393	с	0,04132
в	0,03861	т	0,04999
г	0,01413	у	0,02414
д	0,02493	ф	0,00054
е	0,07344	х	0,00672
ж	0,00939	ц	0,00324
з	0,01307	ч	0,01726
и	0,05965	ш	0,00851
й	0,00870	щ	0,00238
к	0,03790	ы	0,01508
л	0,03993	ь	0,01753
м	0,02562	э	0,00169
н	0,05535	ю	0,00423
о	0,09504	я	0,01461
п	0,02179		0,16070

Табл.2.1 – Частоти букв

Символ	Частота	Символ	Частота	Символ	Частота
аб	0,00059	иц	0,00176	ту	0,00129
ав	0,00285	ич	0,00288	тч	0,00009
аг	0,00041	иш	0,00050	ты	0,00106
ад	0,00144	ищ	0,00012	ть	0,00631
ае	0,00067	ию	0,00029	тю	0,00006
аж	0,00176	ия	0,00144	тя	0,00038
аз	0,00376	и	0,01720	т	0,00414
аи	0,00012	йд	0,00012	уа	0,00003
ай	0,00053	йк	0,00012	уб	0,00067
ак	0,01230	йм	0,00006	ув	0,00059
ал	0,00725	йн	0,00032	уг	0,00132
ам	0,00282	йс	0,00035	уд	0,00211
ан	0,00282	йт	0,00012	уе	0,00018
ап	0,00070	йц	0,00009	уж	0,00205
ар	0,00291	йш	0,00015	уз	0,00021
ас	0,00249	й	0,00787	уй	0,00003
ат	0,00446	ка	0,01147	ук	0,00065
ау	0,00003	кв	0,00035	ул	0,00117
аф	0,00003	ке	0,00067	ум	0,00103
ах	0,00094	кж	0,00003	ун	0,00015
ач	0,00167	ки	0,00581	уп	0,00053
аш	0,00062	кл	0,00035	ур	0,00029
ащ	0,00035	кн	0,00050	ус	0,00109
аю	0,00067	ко	0,00872	ут	0,00097
ая	0,00179	кр	0,00126	ух	0,00065

а	0,01265	кс	0,00003	уч	0,00091
ба	0,00085	кт	0,00056	уш	0,00041
бб	0,00003	ку	0,00202	ущ	0,00012
бе	0,00211	к	0,00546	ую	0,00114
би	0,00041	ла	0,00402	у	0,00643
бк	0,00018	лб	0,00006	фа	0,00015
бл	0,00053	лг	0,00018	фе	0,00006
бн	0,00021	лд	0,00006	фи	0,00009
бо	0,00153	ле	0,00323	фл	0,00006
бп	0,00003	лж	0,00029	фо	0,00003
бр	0,00112	лз	0,00006	фр	0,00009
бс	0,00018	ли	0,00637	фу	0,00006
бу	0,00197	лк	0,00032	фф	0,00003
бх	0,00003	лн	0,00012	ха	0,00067
бщ	0,00018	ло	0,00690	хв	0,00015
бы	0,00440	лп	0,00003	хе	0,00003
бь	0,00009	лс	0,00182	хи	0,00012
бя	0,00015	лт	0,00009	хл	0,00018
б	0,00026	лу	0,00167	хн	0,00012
ва	0,00472	лч	0,00015	хо	0,00191
вв	0,00003	лы	0,00041	хр	0,00003
вд	0,00026	ль	0,00578	хс	0,00003
ве	0,00605	лю	0,00032	хт	0,00006
вз	0,00041	ля	0,00150	ху	0,00009
ви	0,00505	л	0,00701	х	0,00346
вк	0,00029	ма	0,00200	ца	0,00062
вл	0,00044	ме	0,00417	цв	0,00006
вм	0,00018	ми	0,00247	це	0,00076
вн	0,00114	мл	0,00003	ци	0,00006
во	0,00613	мм	0,00006	цк	0,00006
вп	0,00035	мн	0,00112	цм	0,00006
вр	0,00065	мо	0,00326	цо	0,00032
вс	0,00314	мп	0,00015	цу	0,00041
вт	0,00006	му	0,00302	цы	0,00009
ву	0,00050	мц	0,00003	ц	0,00067
вц	0,00003	мч	0,00003	ча	0,00293
вш	0,00141	мы	0,00067	че	0,00335
вы	0,00200	мь	0,00006	чи	0,00226
вь	0,00012	мя	0,00053	чк	0,00038
вя	0,00023	м	0,00880	чн	0,00100
в	0,00546	на	0,00933	чо	0,00015
га	0,00056	нб	0,00003	чр	0,00009
гв	0,00006	нд	0,00026	чт	0,00376
гд	0,00094	не	0,01039	чу	0,00097
ге	0,00029	ни	0,00772	чш	0,00029
ги	0,00085	нк	0,00032	чь	0,00026
гк	0,00009	нн	0,00320	ч	0,00153

гл	0,00079	но	0,01174	ша	0,00070
гн	0,00006	нс	0,00015	шв	0,00012
го	0,00839	нт	0,00050	ше	0,00261
гр	0,00056	ну	0,00255	ши	0,00329
гс	0,00003	нц	0,00015	шк	0,00038
гт	0,00003	нч	0,00006	шл	0,00047
гу	0,00044	нщ	0,00006	шм	0,00021
гч	0,00003	ны	0,00379	шн	0,00015
г	0,00103	нь	0,00132	шо	0,00021
да	0,00531	ню	0,00018	шт	0,00003
дб	0,00009	ня	0,00088	шу	0,00029
дв	0,00076	н	0,00332	шь	0,00018
де	0,00505	об	0,00340	ш	0,00006
дз	0,00003	ов	0,00860	ща	0,00015
ди	0,00156	ог	0,00499	ще	0,00123
дк	0,00050	од	0,00434	щи	0,00067
дл	0,00073	ое	0,00200	щн	0,00006
дм	0,00009	ож	0,00167	щу	0,00009
дн	0,00194	оз	0,00091	ыб	0,00018
до	0,00317	ои	0,00091	ыв	0,00126
др	0,00132	ой	0,00238	ыг	0,00009
дс	0,00029	ок	0,00188	ыд	0,00006
дт	0,00041	ол	0,00514	ые	0,00117
ду	0,00129	ом	0,00619	ыж	0,00006
дц	0,00015	он	0,00534	ыз	0,00009
дч	0,00003	оо	0,00009	ый	0,00156
дш	0,00012	оп	0,00082	ык	0,00056
ды	0,00041	ор	0,00508	ыл	0,00217
дь	0,00070	ос	0,00528	ым	0,00144
дя	0,00009	от	0,00669	ын	0,00023
д	0,00112	оу	0,00021	ып	0,00015
еб	0,00103	ох	0,00032	ыр	0,00023
ев	0,00214	оц	0,00003	ыс	0,00065
ег	0,00361	оч	0,00208	ыт	0,00029
ед	0,00235	ош	0,00082	ых	0,00088
ее	0,00123	ощ	0,00029	ыч	0,00018
еж	0,00094	ою	0,00056	ыш	0,00050
ез	0,00103	оя	0,00065	ыщ	0,00003
еи	0,00006	о	0,02433	ы	0,00320
ей	0,00170	па	0,00106	ьб	0,00006
ек	0,00138	пе	0,00282	ьг	0,00003
ел	0,00754	пи	0,00085	ьд	0,00012
ем	0,00461	пк	0,00018	ье	0,00062
ен	0,00707	пл	0,00085	ьз	0,00023
ео	0,00006	пн	0,00003	ьи	0,00003
еп	0,00109	по	0,00875	ьк	0,00141
ер	0,00537	пр	0,00622	ьм	0,00035

ес	0,00452	пс	0,00003	ьн	0,00161
ет	0,00522	пу	0,00056	ьс	0,00117
еу	0,00012	пы	0,00012	ьт	0,00009
ех	0,00041	пь	0,00003	ьц	0,00009
ец	0,00050	пя	0,00018	ьч	0,00003
еч	0,00106	п	0,00006	ьш	0,00056
еш	0,00056	ра	0,00607	ью	0,00015
ещ	0,00085	рб	0,00012	ья	0,00026
ею	0,00023	рв	0,00018	ь	0,00980
ея	0,00015	рг	0,00029	эй	0,00003
е	0,01828	рд	0,00021	эк	0,00003
жа	0,00112	ре	0,00566	эс	0,00003
жб	0,00003	рж	0,00032	эт	0,00156
жд	0,00038	рз	0,00009	эф	0,00006
же	0,00443	ри	0,00384	юб	0,00023
жи	0,00135	рк	0,00015	юд	0,00026
жк	0,00023	рл	0,00006	юж	0,00003
жн	0,00097	рм	0,00006	юс	0,00006
жр	0,00003	рн	0,00050	ют	0,00029
жс	0,00003	ро	0,00669	юч	0,00012
жу	0,00003	рп	0,00006	ющ	0,00032
жч	0,00003	рр	0,00003	ююю	0,00003
жь	0,00006	рс	0,00023	ю	0,00323
ж	0,00056	рт	0,00141	яб	0,00003
за	0,00472	ру	0,00264	яв	0,00032
зб	0,00015	рх	0,00015	яд	0,00038
зв	0,00070	рч	0,00009	яе	0,00003
зг	0,00012	рш	0,00062	яж	0,00006
зд	0,00070	рщ	0,00003	яз	0,00009
зе	0,00021	ры	0,00138	яи	0,00009
зз	0,00003	рь	0,00044	яй	0,00012
зи	0,00038	рю	0,00009	як	0,00050
зк	0,00006	ря	0,00070	ял	0,00053
зл	0,00018	р	0,00073	ям	0,00053
зм	0,00015	са	0,00164	ян	0,00047
зн	0,00188	св	0,00144	яп	0,00006
зо	0,00073	сг	0,00003	яр	0,00029
зр	0,00012	сд	0,00053	яс	0,00035
зт	0,00003	се	0,00317	ят	0,00161
зу	0,00009	сж	0,00003	ях	0,00012
зч	0,00003	сз	0,00003	яц	0,00006
зы	0,00050	си	0,00117	яч	0,00009
зь	0,00012	ск	0,00308	ящ	0,00009
зя	0,00053	сл	0,00258	яю	0,00006
з	0,00135	см	0,00079	яя	0,00009
иб	0,00082	сн	0,00056	я	0,00883
ив	0,00229	со	0,00258	а	0,00379

иг	0,00035	сп	0,00138	б	0,00610
ид	0,00170	ср	0,00018	в	0,01432
ие	0,00296	сс	0,00065	г	0,00267
иж	0,00018	ст	0,00924	д	0,00831
из	0,00244	су	0,00088	е	0,00434
ии	0,00023	сх	0,00018	ж	0,00211
ий	0,00185	сч	0,00021	з	0,00396
ик	0,00276	сш	0,00006	и	0,01059
ил	0,00496	сы	0,00038	й	0,00003
им	0,00255	сь	0,00299	к	0,01042
ин	0,00428	сю	0,00009	л	0,00332
ип	0,00012	ся	0,00440	м	0,00370
ир	0,00076	с	0,00249	н	0,01705
ис	0,00214	та	0,00616	о	0,00930
ит	0,00364	тв	0,00235	п	0,01623
иф	0,00006	тг	0,00003	р	0,00361
их	0,00173	тд	0,00018	с	0,01605
иц	0,00176	те	0,00428	т	0,00795
ич	0,00288	тз	0,00003	у	0,00434
ил	0,00496	ти	0,00293	ф	0,00035
им	0,00255	тк	0,00035	х	0,00120
ин	0,00428	тл	0,00015	ц	0,00050
ип	0,00012	тн	0,00188	ч	0,00781
ир	0,00076	то	0,01444	ш	0,00173
ис	0,00214	тп	0,00018	щ	0,00012
ит	0,00364	тр	0,00337	э	0,00167
иф	0,00006	тс	0,00109	я	0,00050
их	0,00173	тт	0,00009		

Табл.2.2 – Частоти біграм

3. Значення ентропії

H_1 – ентропія букв алфавіту, H_2 – ентропія біграм, H_3 – ентропія біграм з перетином букв, $H^{(k)}$ – ентропія k-грами.

	Алфавіт з пробілом	Алфавіт без пробілу
H_1	4,36255	3,93869
H_2	7,83589	5,65088
H_3	7,84992	5,66663

Табл.2.3 – Значення ентропії

$1{,}5505 < H^{(10)} < 2{,}4082$
$1{,}2660 < H^{(20)} < 2{,}0180$
$1{,}3224 < H^{(30)} < 2{,}0779$

Табл.2.4 – Оцінка ентропії

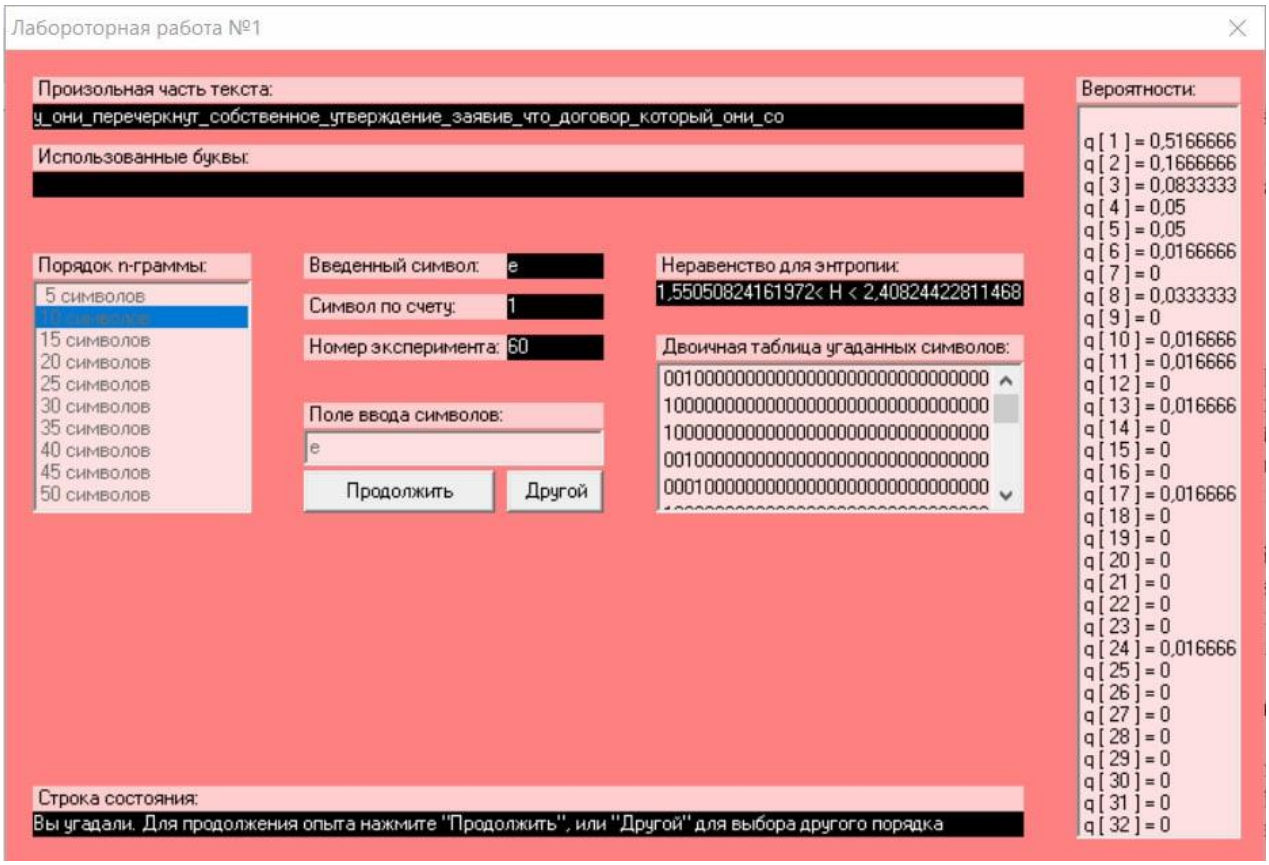

Рис.2.1 – Оцінка $H^{(10)}$

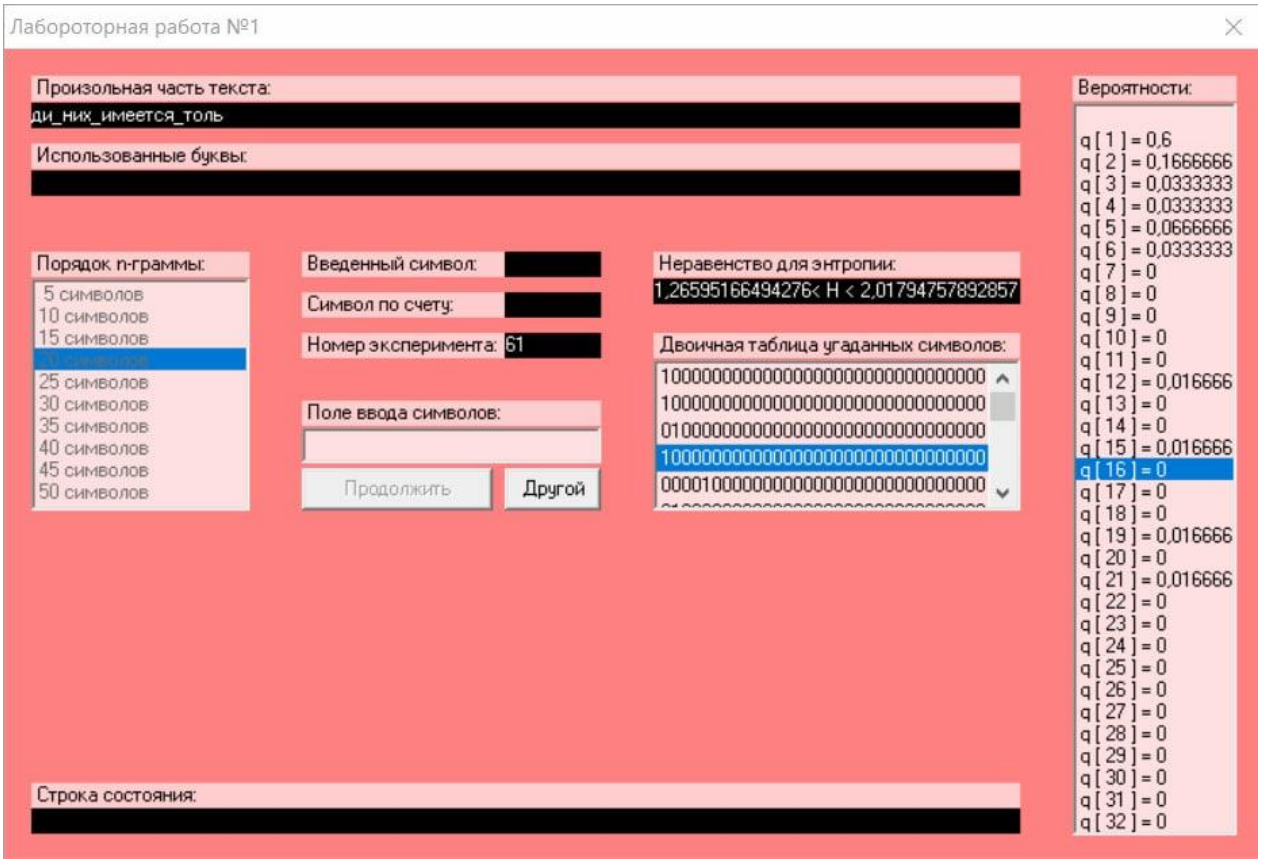

Рис.2.2 – Оцінка $H^{(20)}$

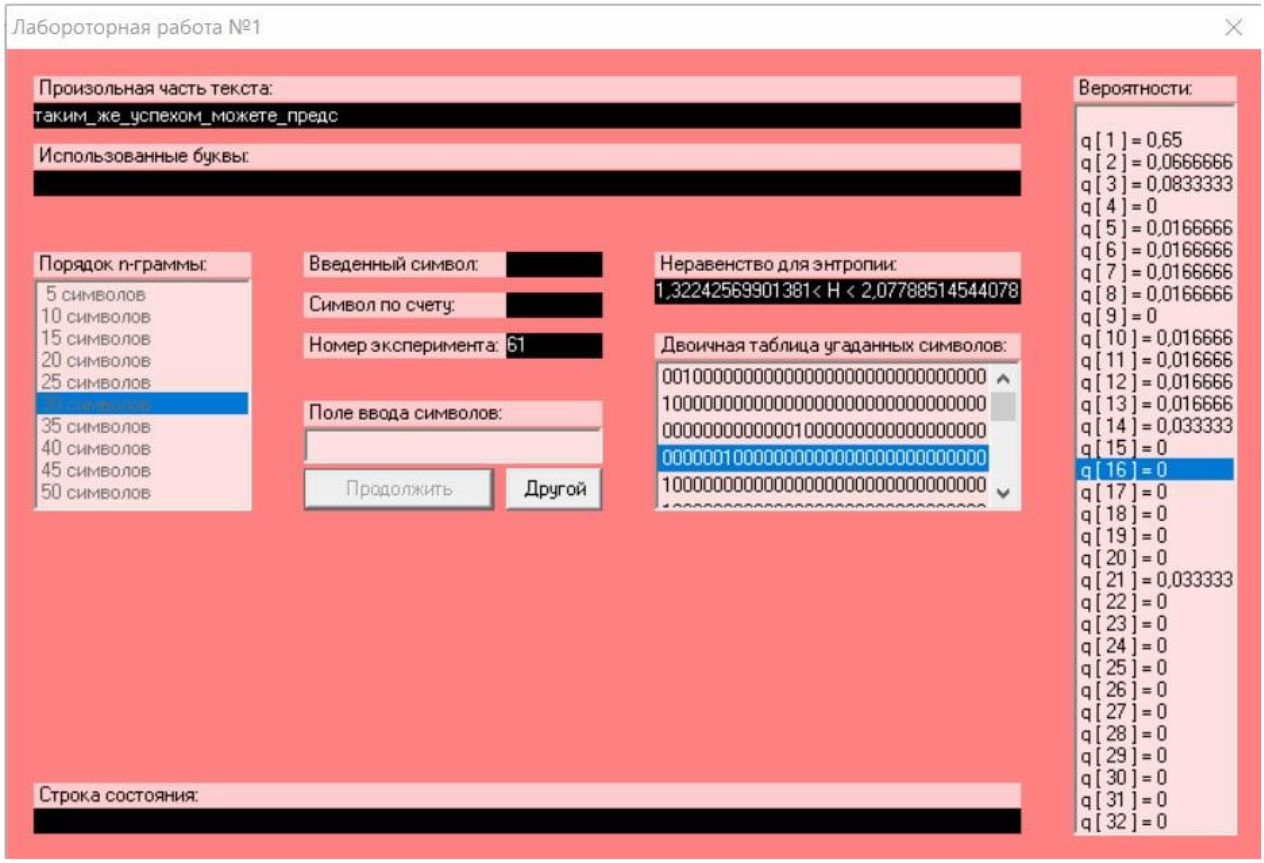

Рис.2.3 – Оцінка $H^{(30)}$

4. Оцінка надлишковості російської мови

$$R = 1 - \frac{H_{real}}{H_{max}} = 1 - \frac{4.36255}{5} = 0.1275$$

ВИСНОВКИ

У даній роботі було обраховано частоти букв і біграм в тексті, а також ентропія за безпосереднім означенням. При підрахунку ентропії біграм як пар літер, що перетинаються, так і пар букв, що не перетинаються, отримані відповіді не сильно відрізнялися одна від одної. Символ пробілу досить часто з’являється у тексті, тому значення ентропії для алфавіту з ним є значно більшим.