Криптографія. Лабораторна робота 1.

Експерементальна оцінка ентропії на символ джерела

Текст та форматування

Як середньостатистичний текст середньостатистичною російською мовою середньостатистичного росіянина ми взяли ~500000 символів (~1Мб) тексту життя терориста та військового злочинця ігоря стрєлкова–гіркіна з його телеграм-каналу.

Зчитуємо текст в змінну-рядок:

```
text = fileread('girkin.txt');
```

Форматуваємо текст (видалення зайвих символів та зайвих пробілів):

```
formated_text = blanks(length(text));

for i = 1:length(text)
    c = text(i);
    if (c <= 'я' && c >= 'а') || c == ' '
        formated_text(i) = c;
    elseif (c <= 'Я' && c >= 'А')
        formated_text(i) = char(c + 0x0020);
    elseif c == 'Ё' || c == 'ё'
        formated_text(i) = 'е';
    else
        formated_text(i) = ' ';
    end
end

mask = (formated_text == ' ');
mask = ~(mask & [0 mask(1:end-1)]);
formated_text2 = formated_text(mask);
```

Оригінальний (файл `girkin.txt`*) та кінцевий форматований текст (змінна* `formated_text2`*) мають бути десь разом з цим файлом.*

Робота з форматованим текстом

Створюємо два відображення типу "символ -> кількість входжень в тексті" та "біграма -> кількість входжень в тексті":

```
bigrams = dictionary(string([]), []);
chars = dictionary(string([]), []);
```

Заповнюємо дані про **окремі символи** :

```
for i = 1:length(formated_text2)
    c1 = formated_text2(i);
    if chars.isKey(c1)
        chars(c1) = chars(c1) + 1;
    else
        chars(c1) = 1;
    end
end
chars;
```

Заповнюємо дані про **біграми** :

```
for i = 1:(length(formated_text2)-1)
    c2 = [formated_text2(i) formated_text2(i+1)];

    if bigrams.isKey(c2)
        bigrams(c2) = bigrams(c2) + 1;
    else
        bigrams(c2) = 1;
    end
end
bigrams;
```

Ці зібрані дані прикладені окремими файлами (див. нижче)

Обчислення ентропій H_1 та H_2

```
t1 = chars.values / sum(chars.values);
H1 = -sum(t1 .* log2(t1))
```

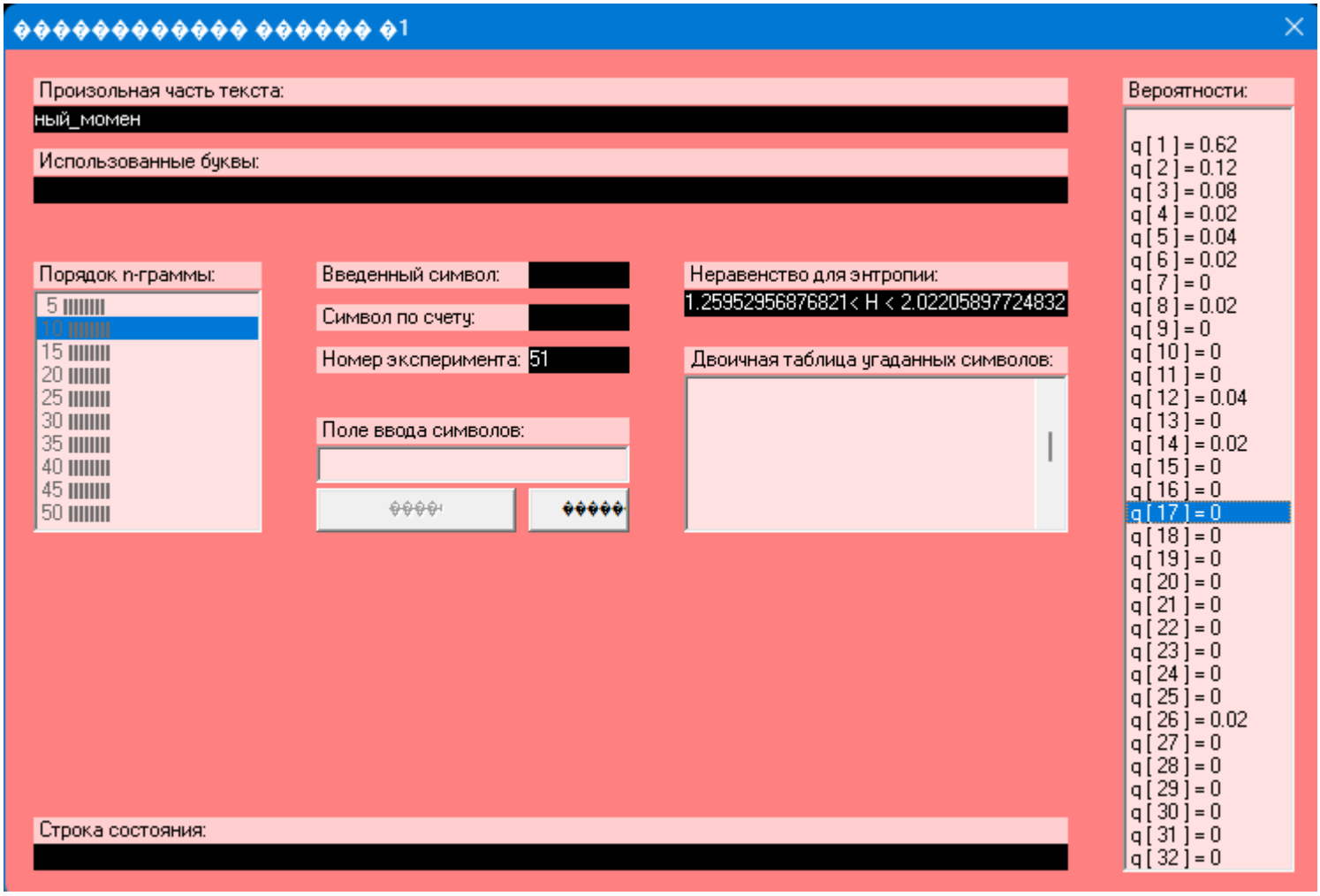
```
H1 = 4.3851
```

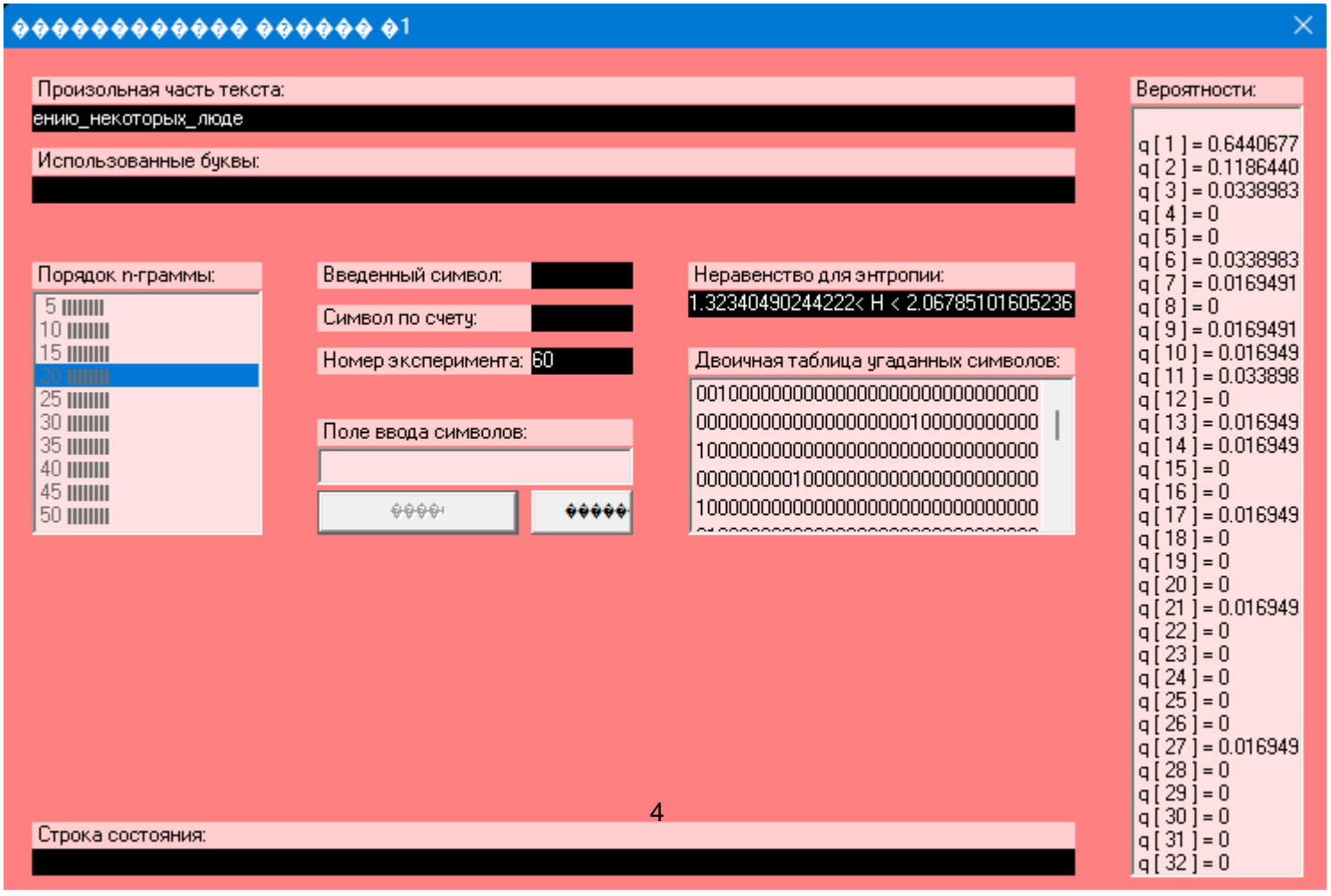
```

t_b = bigrams.values / sum(bigrams.values);
H2 = -sum(t_b .* log2(t_b)) * 0.5
```

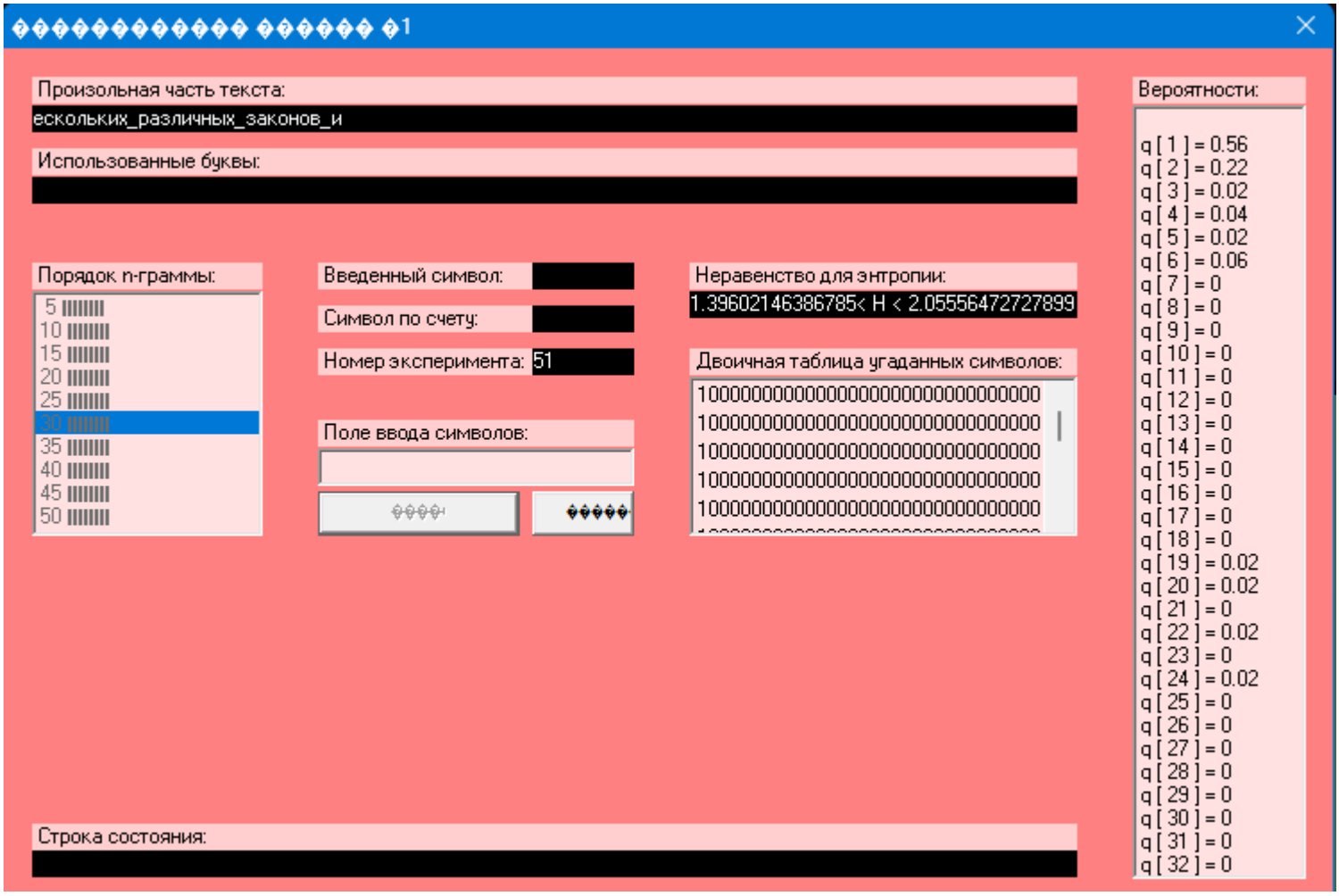
```
H2 = 3.9881
```

Оцінки для $H^{(10)}$, $H^{(20)}$, $H^{(30)}$ та обчислення надлишковості мови

За допомогою `CoolPinkProgram.exe` обчислимо $H^{(10)}$, $H^{(20)}$, $H^{(30)}$:


Произольная часть текста:
ный_момен
Использованные буквы:
Порядок n-граммы:
5
10
15
20
25
30
35
40
45
50
Введенный символ:
Символ по счету:
Номер эксперимента: 51
Поле ввода символов:
Неравенство для энтропии:
1.25952956876821< H < 2.02205897724832
Двоичная таблица угаданных символов:
Строка состояния:
Вероятности:
q [1] = 0.62
q [2] = 0.12
q [3] = 0.08
q [4] = 0.02
q [5] = 0.04
q [6] = 0.02
q [7] = 0
q [8] = 0.02
q [9] = 0
q [10] = 0
q [11] = 0
q [12] = 0.04
q [13] = 0
q [14] = 0.02
q [15] = 0
q [16] = 0
q [17] = 0
q [18] = 0
q [19] = 0
q [20] = 0
q [21] = 0
q [22] = 0
q [23] = 0
q [24] = 0
q [25] = 0
q [26] = 0.02
q [27] = 0
q [28] = 0
q [29] = 0
q [30] = 0
q [31] = 0
q [32] = 0


Произольная часть текста:
ению_некоторых_люде
Использованные буквы:
Порядок n-граммы:
5
10
15
20
25
30
35
40
45
50
Введенный символ:
Символ по счету:
Номер эксперимента: 60
Поле ввода символов:
Неравенство для энтропии:
1.32340490244222< H < 2.06785101605236
Двоичная таблица угаданных символов:
00100000000000000000000000000000
00000000000000000001000000000000
10000000000000000000000000000000
00000000010000000000000000000000
10000000000000000000000000000000
Строка состояния:
Вероятности:
q [1] = 0.6440677
q [2] = 0.1186440
q [3] = 0.0338983
q [4] = 0
q [5] = 0
q [6] = 0.0338983
q [7] = 0.0169491
q [8] = 0
q [9] = 0.0169491
q [10] = 0.016949
q [11] = 0.033898
q [12] = 0
q [13] = 0.016949
q [14] = 0.016949
q [15] = 0
q [16] = 0
q [17] = 0.016949
q [18] = 0
q [19] = 0
q [20] = 0
q [21] = 0.016949
q [22] = 0
q [23] = 0
q [24] = 0
q [25] = 0
q [26] = 0
q [27] = 0.016949
q [28] = 0
q [29] = 0
q [30] = 0
q [31] = 0
q [32] = 0



Отже, маємо такі результати :

$$1.26 \leq H^{(10)} \leq 2.02$$
$$1.32 \leq H^{(20)} \leq 2.07$$
$$1.40 \leq H^{(30)} \leq 2.06$$

Візьмемо $H^{(30)}$ як найкраще наближення для H_∞, тоді обчислимо надлишковість російської мови R за формулою :

$$R = 1 - \frac{H_\infty}{H_0}, \quad \text{де } H_0 = log_2(32) = 5$$

Тоді надлишковість R буде в таких межах :

$$0.58 \leq R \leq 0.72$$

Тобто маємо **надлишковість російської мови в районі 65%.**

Візуалізація

Наступний код потрібен лише для красивих табличок з даними (які також мають бути десь разом з цим файлом) і ніякого іншого корисного навантаження не несе.

(можна було б зробити ці таблички інтерактивними прямо тут, але це занадто складно, а результат не вартує того)

```
keys_c = chars.keys;
values_c = chars.values;
[sortedValues_c, sortInd_c] = sort(values_c);
sortedKeys_c = keys_c(sortInd_c);

%-------------------------------------

svalues1 = zeros(32, 1);

for i = 1:32
    for j = 1:32
        b = string([char(i + 'a' - 1) char(j + 'a' - 1)]);

        if bigrams.isKey(b)
            svalues1(i) = svalues1(i) + bigrams(b);
        end
    end
end

alph = 'абвгдежзийклмнопрстуфхцчшщъыьэюя';

[~, sortInd1] = sort(svalues1);

alph = alph(sortInd1);

bfreq_matrix = zeros(32, 32);

for i = 1:32
    for j = 1:32
        b = string([alph(i) alph(j)]);

        if bigrams.isKey(b)
            bfreq_matrix(i, j) = bigrams(b);
        end
    end
end
```