КРИПТОГРАФІЯ

КОМП’ЮТЕРНИЙ ПРАКТИКУМ №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

ФБ-11 Подолянко Тимофій

Визначення H_1 H_2

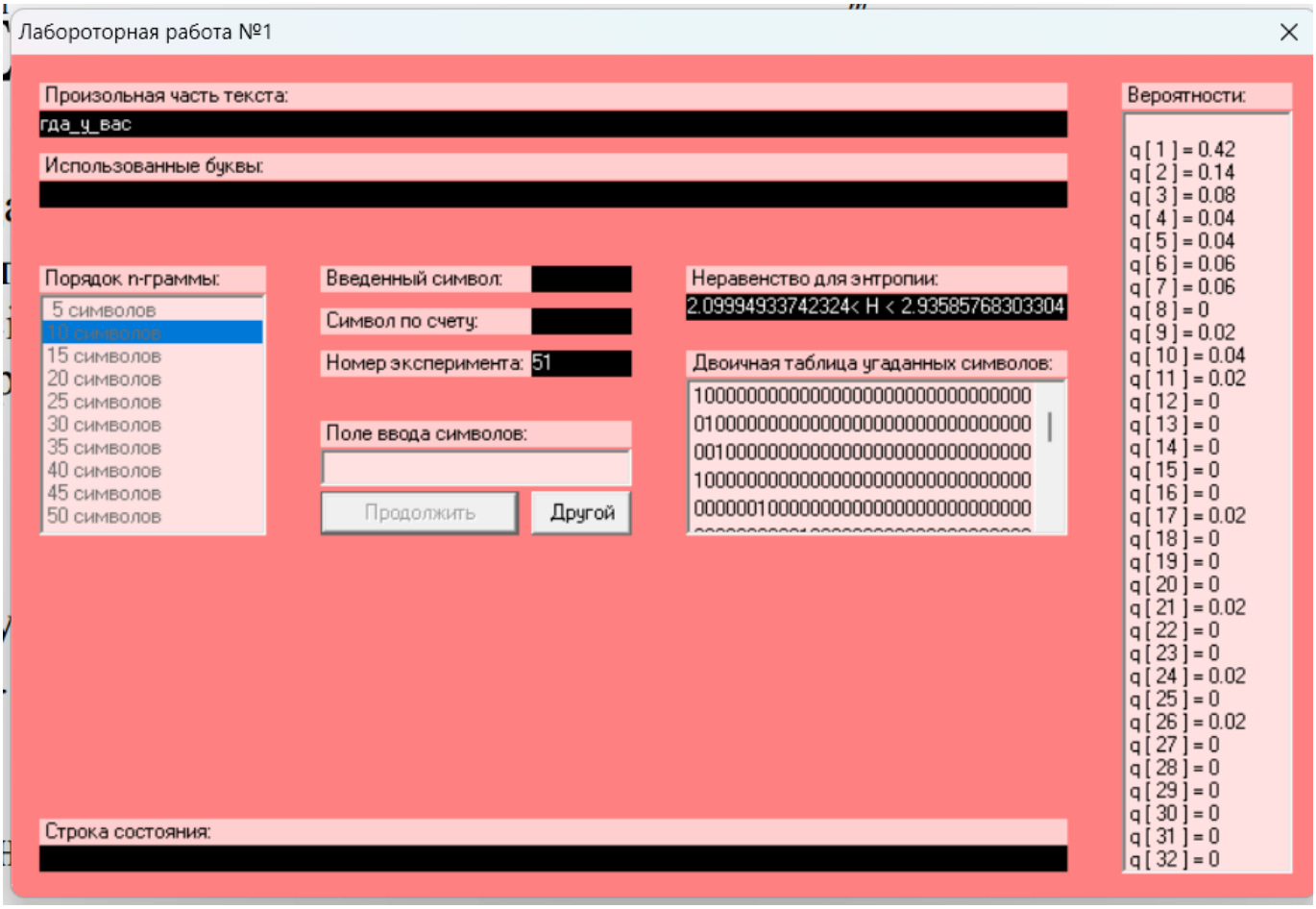
Реалізовано програму, яка проводить нормалізацію тексту (видалення символів, що не належать алфавіту; видалення подвійних пробілів; зміна регістру) та обчислює ентропію букв алфавіту та ентропію на символ біграми у тексті за умов, коли:

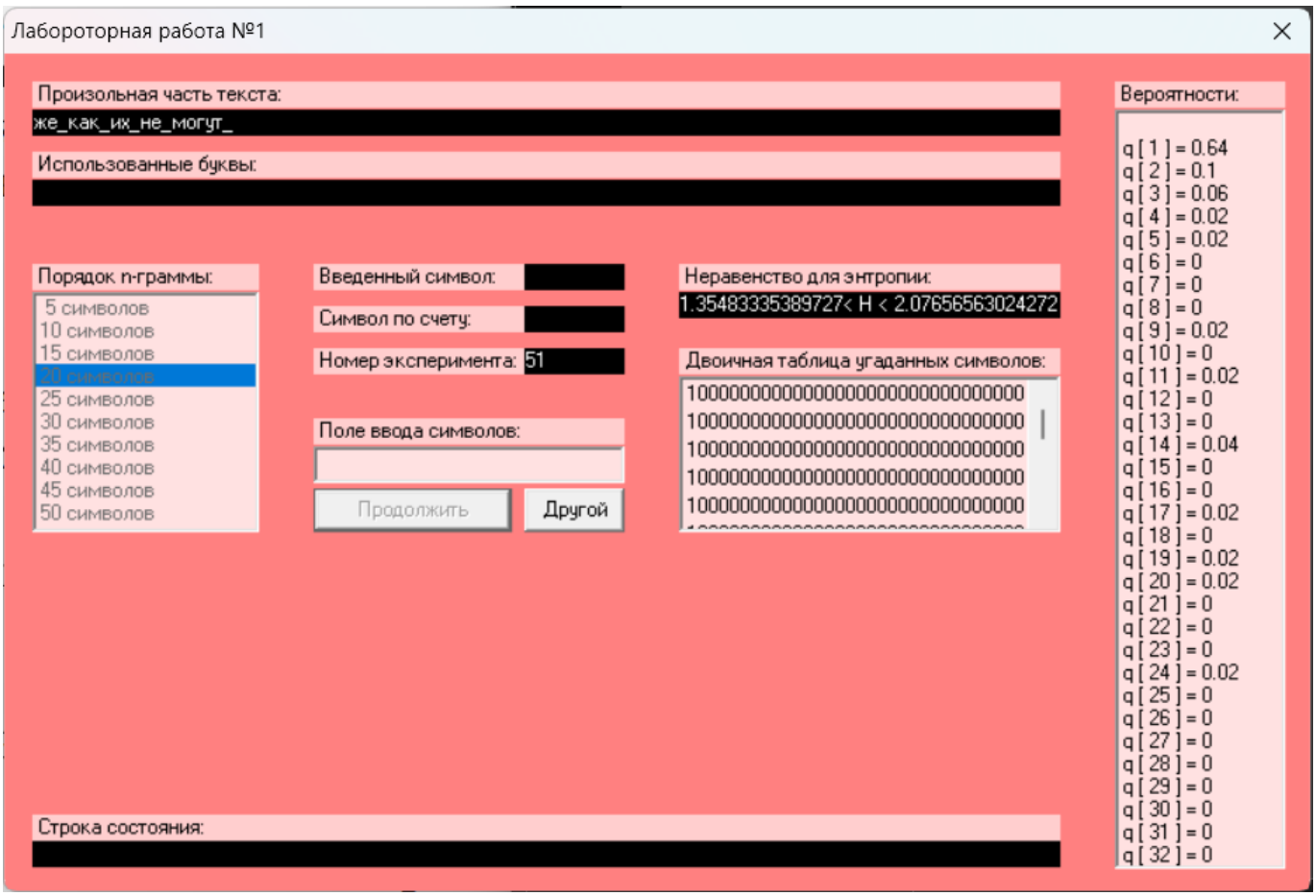
- біграми не перетинаються
- біграми перетинаються
- пробіл входить до алфавіту
- пробіл не входить до алфавіту

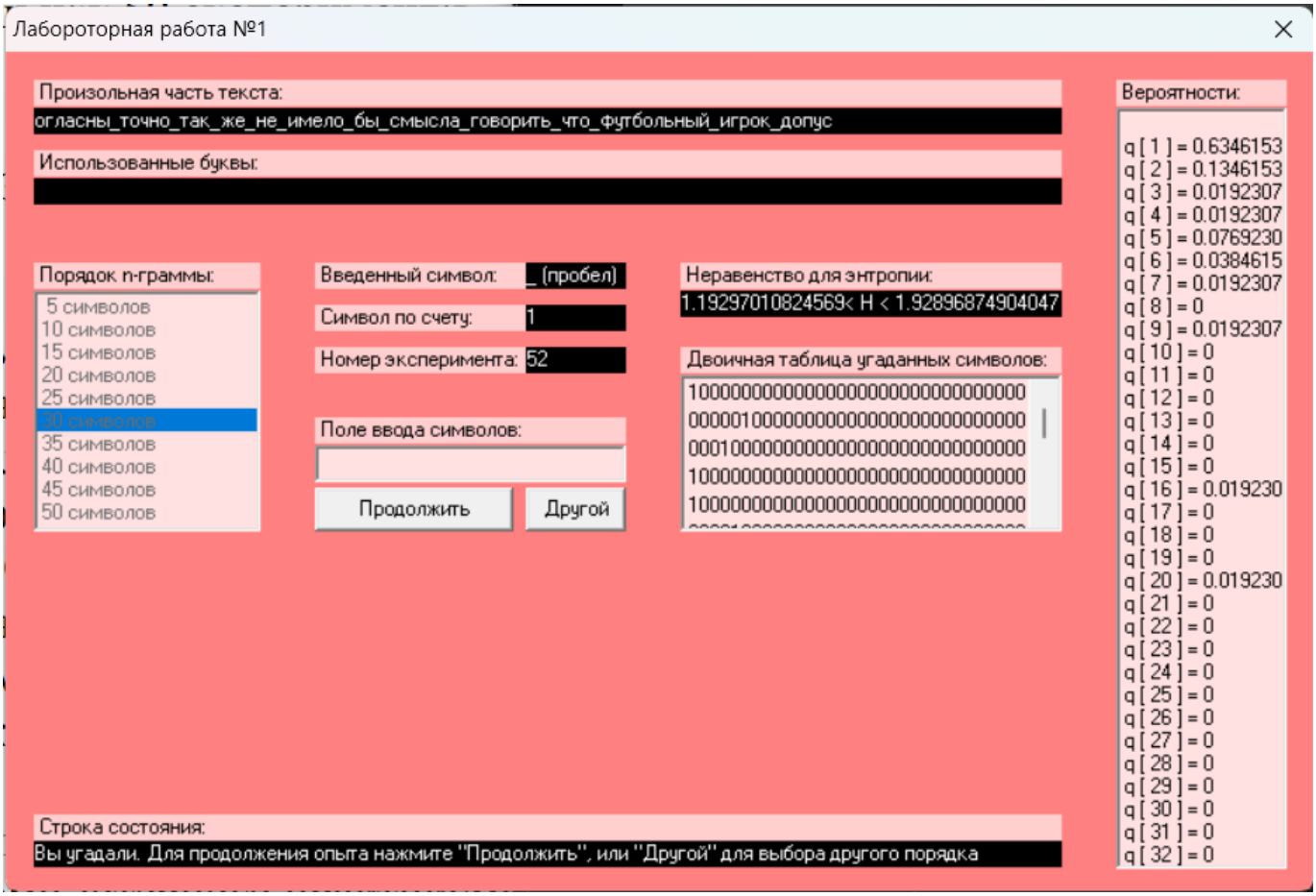
Демонстрація роботи програми

```
PS>cargo run --release .\data\sample.txt
    Finished release [optimized] target(s) in 0.08s
     Running `target\release\podolianko_fb-11_cpl.exe .\data\sample.txt`
Current dir: D:\Documents\UNIV4\Crypto\lab\crypto-23-24\cpl\podolianko_fb-11_cpl
Sample file: .\data\sample.txt

Letter frequencies with whitespaces:

|  :0.1717 | ж:0.0091 | н:0.0522 | ф:0.0096 | ш:0.0136 |
| а:0.0712 | з:0.0150 | о:0.0945 | х:0.0072 | ь:0.0193 |
| б:0.0162 | и:0.0531 | п:0.0228 | ц:0.0030 | э:0.0023 |
| в:0.0360 | й:0.0090 | р:0.0326 | ч:0.0132 | ю:0.0049 |
| г:0.0165 | к:0.0274 | с:0.0412 | ш:0.0069 | я:0.0172 |
| д:0.0272 | л:0.0430 | т:0.0534 | щ:0.0022 | ё:0.0001 |
| е:0.0693 | м:0.0247 | у:0.0244 | ъ:0.0002 |
Entropy: 4.352013962389525
Redundancy: 0.1445610832052681128

Letter frequencies without whitespaces:

| а:0.0860 | з:0.0181 | о:0.1141 | х:0.0087 | ь:0.0233 |
| б:0.0195 | и:0.0642 | п:0.0265 | ц:0.0036 | э:0.0028 |
| в:0.0434 | й:0.0109 | р:0.0393 | ч:0.0160 | ю:0.0059 |
| г:0.0199 | к:0.0331 | с:0.0498 | ш:0.0083 | я:0.0208 |
| д:0.0328 | л:0.0519 | т:0.0644 | щ:0.0026 | ё:0.0001 |
| е:0.0836 | м:0.0298 | у:0.0294 | ъ:0.0002 |
| ж:0.0110 | н:0.0630 | ф:0.0008 | ы:0.0164 |
Entropy: 4.455304233859921
Redundancy: 0.11678109829639027

Bigram frequencies overlapping, with whitespace:

|      |      | а | б | в | г | д | е | ж | з | и | й | к | л | м | н | о | п | р | с | т | у | ф | х | ц | ч | ш | щ | ъ | ы | ь | э | ю | я | ё |
|------|------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|      |  _   |0.00292|0.00662|0.01468|0.00431|0.00851|0.00428|0.00265|0.00561|0.01106|  _   |0.00833|0.00310|0.00529|0.01777|0.01436|0.01528|0.00354|0.01595|0.00955|0.00457|0.00029|0.00158|0.00037|0.00542|0.00120|0.00010|  _   |  _   |  _   |0.00225|0.00003|0.00205|  _   |
|  а   |0.02022|  _   |0.00071|0.00244|0.00077|0.00209|0.00158|0.00129|0.00355|0.00010|0.00074|0.00537|0.00830|0.00297|0.00340|0.00001|0.00071|0.00294|0.00345|0.00436|0.00007|0.00014|0.00160|0.00005|0.00095|0.00067|0.00018|  _   |  _   |  _   |0.00000|0.00087|0.00174|  _   |
|  б   |0.00068|0.00118|0.00001|0.00012|0.00000|0.00002|0.00208|0.00002|  _   |0.00077|  _   |0.00019|0.00226|0.00005|0.00020|0.00223|  _   |0.00118|0.00008|0.00001|0.00144|  _   |0.00004|0.00000|0.00001|0.00001|0.00010|0.00008|0.00274|0.00005|  _   |0.00001|0.00059|  _   |
|  в   |0.00656|0.00531|0.00001|0.00082|0.00056|0.00437|  _   |0.00075|0.00257|  _   |0.00225|0.00662|0.00009|0.00107|0.00646|0.00018|0.00034|0.00270|0.00033|0.00665|  _   |0.00003|0.00003|0.00008|0.00027|0.00002|0.00006|0.00222|0.00025|  _   |0.00000|0.00019|  _   |
|  г   |0.00099|0.00135|  _   |0.00001|  _   |0.00138|0.00028|  _   |  _   |  _   |0.00076|  _   |0.00013|0.00176|0.00001|0.00023|0.00811|  _   |0.00080|0.00004|0.00001|0.00064|  _   |  _   |  _   |  _   |0.00001|0.00006|  _   |  _   |  _   |  _   |  _   |0.00000|  _   |  _   |
|  д   |0.00100|0.00525|0.00002|0.00100|0.00001|0.00002|0.00483|0.00001|0.00001|0.00215|  _   |0.00026|0.00046|0.00004|0.00161|0.00389|0.00009|0.00179|0.00021|0.00031|0.00195|  _   |0.00003|0.00032|0.00003|0.00004|  _   |0.00001|0.00062|0.00081|  _   |0.00001|0.00039|  _   |
|  е   |0.01832|0.00086|0.00145|0.00165|0.00321|0.00246|0.00138|0.00090|0.00130|0.00017|0.00218|0.00094|0.00520|0.00368|0.00571|0.00014|0.00098|0.00475|0.00353|0.00645|0.00009|0.00000|0.00058|0.00034|0.00110|0.00103|0.00072|  _   |  _   |  _   |  _   |0.00017|0.00025|  _   |
|  ж   |0.00052|0.00124|0.00006|  _   |0.00001|0.00008|0.00324|0.00002|  _   |0.00191|  _   |0.00008|0.00001|0.00002|0.00058|0.00006|  _   |  _   |  _   |0.00001|  _   |0.00023|  _   |  _   |  _   |  _   |0.00004|  _   |  _   |  _   |  _   |0.00003|  _   |  _   |  _   |  _   |
|  з   |0.00118|0.00638|0.00017|0.00068|0.00045|0.00096|0.00025|0.00011|0.00002|0.00041|  _   |0.00008|0.00019|0.00021|0.00197|0.00034|  _   |0.00023|0.00002|0.00002|0.00021|  _   |  _   |  _   |0.00001|0.00003|0.00000|  _   |0.00001|0.00041|0.00011|  _   |0.00000|0.00053|  _   |
|  и   |0.01820|0.00003|0.00074|0.00234|0.00046|0.00131|0.00171|0.00024|0.00180|0.00034|0.00074|0.00140|0.00630|0.00195|0.00267|0.00066|0.00023|0.00073|0.00226|0.00429|0.00000|0.00001|0.00123|0.00072|0.00132|0.00074|0.00014|  _   |  _   |  _   |  _   |0.00023|0.00097|  _   |
|  й   |0.00718|  _   |  _   |  _   |  _   |0.00027|0.00011|  _   |0.00000|  _   |  _   |0.00022|0.00002|0.00004|0.00032|  _   |  _   |0.00009|0.00039|0.00026|  _   |  _   |  _   |0.00002|0.00011|0.00045|0.00000|  _   |  _   |  _   |  _   |  _   |  _   |  _   |  _   |  _   |
|  к   |0.00496|0.00778|  _   |0.00022|  _   |  _   |0.00048|0.00001|0.00000|0.00229|  _   |0.00001|0.00048|0.00000|0.00043|0.00692|  _   |0.00129|0.00013|0.00048|0.00190|  _   |0.00000|0.00001|  _   |0.00000|  _   |  _   |  _   |  _   |  _   |  _   |  _   |  _   |  _   |
|  л   |0.00737|0.00668|0.00003|0.00002|0.00022|0.00002|0.00356|0.00031|0.00002|0.00575|  _   |0.00037|0.00007|0.00001|0.00031|0.00659|0.00005|  _   |0.00118|0.00007|0.00146|0.00000|0.00000|  _   |0.00026|0.00001|0.00000|  _   |0.00079|0.00490|  _   |0.00114|0.00174|0.00002|
|  м   |0.00712|0.00249|0.00001|0.00000|0.00004|  _   |0.00330|  _   |0.00000|0.00230|  _   |0.00008|0.00016|0.00002|0.00127|0.00439|0.00007|0.00006|0.00007|0.00062|0.00197|0.00000|  _   |0.00004|0.00012|0.00000|0.00000|  _   |0.00072|0.00009|  _   |  _   |0.00029|  _   |
|  н   |0.00420|0.01069|0.00003|0.00001|0.00006|0.00033|0.01110|0.00001|0.00001|0.00711|  _   |0.00024|0.00000|  _   |0.00185|0.00793|  _   |0.00006|0.00024|0.00043|0.00233|0.00001|0.00000|0.00017|0.00016|  _   |0.00014|  _   |0.00234|0.00130|  _   |0.00025|0.00134|  _   |
|  о   |0.02109|  _   |0.00457|0.00728|0.00428|0.00419|0.00175|0.00177|0.00114|0.00058|0.00377|0.00180|0.00610|0.00604|0.00715|0.00017|0.00144|0.00497|0.00588|0.00651|0.00005|0.00011|0.00053|0.00008|0.00140|0.00094|0.00016|  _   |  _   |  _   |0.00005|0.00029|0.00042|  _   |
|  п   |0.00086|0.00110|  _   |  _   |  _   |  _   |0.00219|  _   |  _   |0.00114|  _   |0.00008|0.00074|  _   |0.00011|0.00874|0.00002|0.00607|0.00000|0.00011|0.00060|  _   |0.00000|0.00001|0.00002|0.00003|0.00000|  _   |0.00028|0.00007|  _   |  _   |0.00058|  _   |
|  р   |0.00105|0.00628|0.00007|0.00026|0.00018|0.00044|0.00529|0.00021|0.00005|0.00449|  _   |0.00027|0.00005|0.00012|0.00053|0.00654|0.00010|0.00001|0.00015|0.00058|0.00296|0.00001|0.00009|0.00002|0.00012|0.00008|0.00002|  _   |0.00105|0.00054|  _   |0.00018|0.00083|  _   |
|  с   |0.00316|0.00138|0.00006|0.00119|0.00002|0.00024|0.00355|0.00002|0.00001|0.00166|  _   |0.00159|0.00275|0.00008|0.00078|0.00193|0.00186|0.00017|0.00039|0.00081|0.00057|0.00003|0.00014|0.00004|0.00044|0.00003|  _   |0.00006|0.00011|0.00300|  _   |0.00016|0.00312|0.00006|
|  т   |0.00787|0.00572|0.00001|0.00204|0.00001|0.00012|0.00467|0.00000|0.00001|0.00315|  _   |0.00062|0.00031|0.00002|0.00076|0.01277|0.00006|0.00266|0.00150|0.00010|0.00123|0.00000|0.00001|0.00013|0.00023|0.00000|0.00001|0.00001|0.00191|0.00692|  _   |0.00006|0.00047|  _   |
|  у   |0.00682|0.00001|0.00057|0.00056|0.00165|0.00215|0.00026|0.00134|0.00024|0.00001|0.00021|0.00084|0.00113|0.00154|0.00015|0.00001|0.00062|0.00038|0.00100|0.00155|  _   |0.00002|0.00049|0.00000|0.00095|0.00079|0.00023|  _   |  _   |  _   |  _   |0.00080|0.00065|  _   |
|  ф   |0.00002|0.00009|  _   |  _   |  _   |  _   |0.00013|  _   |  _   |  _   |0.00008|  _   |  _   |0.00002|0.00000|0.00000|0.00006|  _   |0.00011|0.00001|0.00000|0.00003|  _   |  _   |  _   |  _   |0.00000|0.00000|  _   |  _   |0.00001|0.00008|  _   |  _   |0.00000|  _   |
|  х   |0.00245|0.00136|  _   |0.00018|  _   |  _   |0.00006|  _   |  _   |0.00014|  _   |  _   |0.00018|0.00002|0.00033|0.00231|  _   |0.00010|0.00002|0.00001|0.00010|  _   |  _   |  _   |0.00000|0.00001|  _   |0.00008|  _   |  _   |0.00001|  _   |  _   |  _   |  _   |  _   |
|  ц   |0.00065|0.00069|  _   |0.00008|  _   |  _   |0.00076|  _   |  _   |0.00011|  _   |0.00002|  _   |0.00000|  _   |0.00022|  _   |  _   |0.00000|  _   |0.00023|  _   |  _   |0.00000|  _   |  _   |  _   |  _   |0.00019|  _   |  _   |  _   |  _   |  _   |  _   |
|  ч   |0.00053|0.00232|  _   |0.00001|  _   |  _   |0.00298|  _   |  _   |0.00174|  _   |0.00026|0.00002|0.00001|0.00063|0.00068|  _   |0.00003|  _   |0.00364|0.00066|  _   |  _   |  _   |  _   |  _   |0.00011|  _   |  _   |  _   |0.00021|  _   |  _   |  _   |  _   |  _   |
|  ш   |0.00008|0.00084|  _   |0.00002|  _   |  _   |0.00164|  _   |  _   |0.00126|  _   |0.00036|0.00050|0.00001|0.00032|0.00019|0.00000|  _   |  _   |0.00045|0.00028|  _   |  _   |  _   |  _   |  _   |  _   |  _   |  _   |0.00093|  _   |  _   |  _   |  _   |  _   |
|  щ   |0.00001|0.00030|  _   |  _   |  _   |  _   |0.00116|  _   |  _   |0.00058|  _   |  _   |  _   |  _   |0.00003|0.00000|  _   |0.00000|  _   |  _   |0.00010|  _   |  _   |  _   |  _   |  _   |  _   |  _   |0.00003|  _   |  _   |  _   |  _   |  _   |
|  ъ   |  _   |  _   |  _   |  _   |  _   |  _   |0.00010|  _   |  _   |  _   |  _   |  _   |  _   |  _   |  _   |  _   |  _   |  _   |  _   |  _   |  _   |  _   |  _   |  _   |  _   |  _   |  _   |  _   |  _   |  _   |  _   |0.00008|  _   |
|  ы   |0.00445|  _   |0.00022|0.00096|0.00018|0.00020|0.00077|0.00001|0.00005|0.00000|0.00114|0.00020|0.00149|0.00085|0.00023|  _   |0.00019|0.00022|0.00056|0.00066|0.00002|  _   |0.00066|0.00001|0.00014|0.00043|0.00001|  _   |  _   |  _   |  _   |  _   |0.00000|  _   |
|  ь   |0.01137|  _   |0.00015|0.00001|0.00060|0.00003|0.00087|0.00000|0.00013|0.00051|  _   |0.00118|  _   |0.00031|0.00092|0.00001|0.00090|  _   |0.00007|0.00007|  _   |0.00000|  _   |0.00052|0.00006|0.00032|0.00000|  _   |  _   |  _   |  _   |  _   |0.00054|0.00056|0.00001|
|  э   |0.00002|  _   |  _   |  _   |0.00001|  _   |  _   |  _   |0.00001|  _   |0.00001|0.00005|0.00001|0.00001|0.00008|  _   |0.00001|0.00000|0.00000|0.00215|  _   |0.00000|0.00001|  _   |  _   |  _   |  _   |  _   |  _   |  _   |  _   |  _   |  _   |  _   |
|  ю   |0.00275|  _   |0.00065|  _   |0.00000|0.00023|0.00000|0.00001|0.00000|  _   |  _   |0.00000|0.00001|0.00005|0.00003|  _   |  _   |0.00004|0.00012|0.00045|  _   |0.00000|0.00003|0.00000|0.00009|0.00011|0.00024|  _   |  _   |  _   |  _   |  _   |0.00002|  _   |  _   |
|  я   |0.01046|  _   |0.00003|0.00027|0.00010|0.00084|0.00012|0.00015|0.00022|0.00003|0.00023|0.00019|0.00071|0.00031|0.00058|  _   |0.00005|0.00008|0.00053|0.00170|  _   |  _   |0.00013|0.00006|0.00015|0.00000|0.00012|  _   |  _   |  _   |  _   |  _   |0.00009|0.00008|  _   |
|  ё   |0.00006|  _   |  _   |0.00001|  _   |  _   |  _   |  _   |  _   |  _   |  _   |  _   |  _   |  _   |  _   |  _   |  _   |  _   |  _   |0.00000|  _   |0.00000|  _   |  _   |  _   |  _   |  _   |  _   |  _   |  _   |  _   |  _   |  _   |  _   |
Entropy per character: 3.944880148771585
Redundancy: 0.22460207144323052
```

Оцінка значень $H^{(10)}$, $H^{(20)}$, $H^{(30)}$


Лабороторная работа №1
Произольная часть текста:
гда_у_вас
Использованные буквы:
Порядок n-граммы:
5 символов
10 символов
15 символов
20 символов
25 символов
30 символов
35 символов
40 символов
45 символов
50 символов
Введенный символ:
Символ по счету:
Номер эксперимента: 51
Поле ввода символов:
Продолжить
Другой
Неравенство для энтропии:
2.09994933742324< H < 2.93585768303304
Двоичная таблица угаданных символов:
Вероятности:
q [1] = 0.42
q [2] = 0.14
q [3] = 0.08
q [4] = 0.04
q [5] = 0.04
q [6] = 0.06
q [7] = 0.06
q [8] = 0
q [9] = 0.02
q [10] = 0.04
q [11] = 0.02
q [12] = 0
q [13] = 0
q [14] = 0
q [15] = 0
q [16] = 0
q [17] = 0.02
q [18] = 0
q [19] = 0
q [20] = 0
q [21] = 0.02
q [22] = 0
q [23] = 0
q [24] = 0.02
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
Строка состояния:


Лабороторная работа №1
Произольная часть текста:
же_как_их_не_могут_
Использованные буквы:
Порядок n-граммы:
5 символов
10 символов
15 символов
20 символов
25 символов
30 символов
35 символов
40 символов
45 символов
50 символов
Введенный символ:
Символ по счету:
Номер эксперимента: 51
Поле ввода символов:
Продолжить
Другой
Неравенство для энтропии:
1.35483335389727< H < 2.07656563024272
Двоичная таблица угаданных символов:
Вероятности:
q [1] = 0.64
q [2] = 0.1
q [3] = 0.06
q [4] = 0.02
q [5] = 0.02
q [6] = 0
q [7] = 0
q [8] = 0
q [9] = 0.02
q [10] = 0
q [11] = 0.02
q [12] = 0
q [13] = 0
q [14] = 0.04
q [15] = 0
q [16] = 0
q [17] = 0.02
q [18] = 0
q [19] = 0.02
q [20] = 0.02
q [21] = 0
q [22] = 0
q [23] = 0
q [24] = 0.02
q [25] = 0
q [26] = 0
q [27] = 0
q [28] = 0
q [29] = 0
q [30] = 0
q [31] = 0
q [32] = 0
Строка состояния:



Результати оцінки

$R = H_\infty / H_0$, де H_∞ — ентропія джерела, або, в нашому випадку, її наближення ч

$$H_0 = \log_2 m, \text{ де — кількість символів алфавіту}$$

Обчислимо R для різних отриманих значень ентропіїї з округленням до третьої цифри після коми.

Для ентропій, обчислених з частот букв та біграм тексту:

- коли пробіл входить до алфавіту ($m = 34$):

		R
H_1	4.352013962389526	0.14456103205268112
H_2 з перетинами	3.9448081487150466	0.2246020714432304
H_2 без перетинів	3.944643848640013	0.22463436653415614

- коли пробіл **не** входить до алфавіту ($m = 33$):

		R
H_1	4.45530423385992	0.1167810982963905
H_2 з перетинами	4.13811270269071	0.1796611040342353
H_2 без перетинів	4.136940363004921	0.17989350849313612

Для умовних ентропій, отриманих за допомогою `CollPinkProgram` ($m = 32$):

		R
$H^{(10)}$	$2.09994933742324 < H < 2.93585768303304$	$0.4128284633933921 < R < 0.5800101325153519$
$H^{(20)}$	$1.35483335389727 < H < 2.07656563024272$	$0.584686873951456 < R < 0.729033329220546$
$H^{(30)}$	$1.19297010824569 < H < 1.92896874904047$	$0.614206250191906 < R < 0.761405978350862$

Висновки

Результати проведених експериментів показують значну надлишковість природної мови ($R_{max} = 0.761405978350862$), однак досягти відповідного (чи хоча б порівняного) ущільнення тексту на практиці дуже складно, оскільки наведене значення надлишковості враховує безліч ваємозв'язків та закономірностей у природній мові.