Комп’ютерний практикум №1

**Експериментальна оцінка ентропії на символ джерела відкритого тексту**

Виконала Годлевська Аліса ФБ-11

**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.

2. За допомогою програми CoolPinkProgram оцінити значення (10) H , (20) H , (30) H .

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

**Хід роботи**

Спочатку я працювала із текстом gonegirl.txt. Я прибрала усі символи, крім російських літер, замінила ъ на ь, ё на е. Відредагований текст з пробілами називається edited.txt, без пробілів – edited\_no\_spaces.txt.

Далі я створила функції для підрахунку частот монограм та біграм. Для цього треба просто поділити кількість появлень монограми/біграми на загальну кількість монограм/біграм. Звідси я вже могла отримати значення для ентропій H1 та H2. (*Усі результати вносяться в документ Excel – output\_file.xlsx*.)

Для монограм використовувала таку формулу:

Для біграм таку:

А потім вже отримала значення для надлишковості:

Таблиця 1. Монограми з пробілом

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Монограма** | **Частота** | **Ентропія** | **Загальна ентропія** | **Надлишковість** |
|  | 0,1624 | 0,128746 | 4,395512 | 0,120898 |
| о | 0,0899 | 0,100817 |  |  |
| а | 0,0702 | 0,024446 |  |  |
| е | 0,0686 | 0,169617 |  |  |
| н | 0,0569 | 0,265183 |  |  |
| и | 0,0555 | 0,138424 |  |  |
| т | 0,0518 | 0,085746 |  |  |
| с | 0,0423 | 0,221233 |  |  |
| л | 0,0385 | 0,235306 |  |  |
| в | 0,0351 | 0,036346 |  |  |
| р | 0,0349 | 0,312451 |  |  |
| м | 0,0295 | 0,025834 |  |  |
| к | 0,0287 | 0,090421 |  |  |
| д | 0,0264 | 0,149953 |  |  |
| п | 0,0244 | 0,180911 |  |  |
| у | 0,0239 | 0,045468 |  |  |
| я | 0,0219 | 0,081931 |  |  |
| ь | 0,0172 | 0,081449 |  |  |
| ы | 0,0161 | 0,014071 |  |  |
| б | 0,0149 | 0,063291 |  |  |
| з | 0,0139 | 0,269033 |  |  |
| г | 0,0131 | 0,095904 |  |  |
| ч | 0,013 | 0,425874 |  |  |
| ж | 0,0096 | 0,193023 |  |  |
| й | 0,0094 | 0,147025 |  |  |
| ш | 0,0067 | 0,231511 |  |  |
| х | 0,0067 | 0,168938 |  |  |
| ю | 0,0062 | 0,13071 |  |  |
| э | 0,0047 | 0,048385 |  |  |
| щ | 0,0031 | 0,120733 |  |  |
| ц | 0,0029 | 0,048385 |  |  |
| ф | 0,0015 | 0,064346 |  |  |

Таблиця 2. Монограми без пробілів

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Монограма** | **Частота** | **Ентропія** | **Загальна ентропія** | **Надлишковість** |
| о | 0,1073 | 0,204042 | 4,484176 | 0,094873 |
| а | 0,0838 | 0,052379 |  |  |
| е | 0,0818 | 0,146656 |  |  |
| н | 0,0679 | 0,093636 |  |  |
| и | 0,0663 | 0,115385 |  |  |
| т | 0,0618 | 0,027881 |  |  |
| с | 0,0505 | 0,191772 |  |  |
| л | 0,0459 | 0,295442 |  |  |
| в | 0,0419 | 0,093636 |  |  |
| р | 0,0417 | 0,055726 |  |  |
| м | 0,0352 | 0,016412 |  |  |
| к | 0,0343 | 0,157138 |  |  |
| д | 0,0315 | 0,09815 |  |  |
| п | 0,0291 | 0,248204 |  |  |
| у | 0,0286 | 0,07258 |  |  |
| я | 0,0262 | 0,263482 |  |  |
| ь | 0,0206 | 0,299745 |  |  |
| ы | 0,0192 | 0,109493 |  |  |
| б | 0,0178 | 0,217532 |  |  |
| з | 0,0166 | 0,074086 |  |  |
| г | 0,0156 | 0,166892 |  |  |
| ч | 0,0156 | 0,259554 |  |  |
| ж | 0,0115 | 0,042492 |  |  |
| й | 0,0112 | 0,191145 |  |  |
| ш | 0,008 | 0,345536 |  |  |
| х | 0,008 | 0,148493 |  |  |
| ю | 0,0074 | 0,02989 |  |  |
| э | 0,0057 | 0,103453 |  |  |
| щ | 0,0037 | 0,137662 |  |  |
| ц | 0,0034 | 0,055726 |  |  |
| ф | 0,0018 | 0,169955 |  |  |

Далі я не буду всі дані таблиць заносити в протокол, повний варіант буде в документі excel.

Таблиця 3. Біграми з перетином та з пробілом.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Біграма** | **Частота** | **Ентропія** | **Загальна ентропія** | **Надлишковість** |
| а | 0,018884 | 0,06598 | 3,993261 | 0,201348 |
| о | 0,018527 | 0,049977 |  |  |
| е | 0,017579 | 0,047137 |  |  |
| п | 0,017011 | 0,01871 |  |  |
| и | 0,017 | 0,02335 |  |  |
| н | 0,01659 | 0,00348 |  |  |
| я | 0,01525 | 0,078374 |  |  |
| в | 0,014541 | 0,036541 |  |  |
| с | 0,014426 | 0,052204 |  |  |
| то | 0,012367 | 0,108143 |  |  |
| на | 0,011282 | 0,021278 |  |  |
| ь | 0,010904 | 0,052035 |  |  |
| о | 0,009912 | 0,001329 |  |  |
| по | 0,009838 | 0,006459 |  |  |
| не | 0,009429 | 0,041968 |  |  |
| но | 0,009377 | 0,001505 |  |  |
| ст | 0,009371 | 0,035917 |  |  |
| м | 0,009288 | 0,040476 |  |  |
| к | 0,008439 | 0,062698 |  |  |
| д | 0,008061 | 0,017555 |  |  |
| и | 0,007903 | 0,06317 |  |  |
| ен | 0,007782 | 0,016917 |  |  |
| ни | 0,007716 | 0,029863 |  |  |
| й | 0,007523 | 0,030485 |  |  |
| ко | 0,007395 | 0,063138 |  |  |
| ра | 0,007369 | 0,018022 |  |  |
| ла | 0,007339 | 0,043523 |  |  |
| ал | 0,007184 | 0,09998 |  |  |
| ка | 0,007069 | 0,048125 |  |  |
| ро | 0,006995 | 0,028218 |  |  |
| от | 0,006977 | 0,003623 |  |  |
| м | 0,006954 | 0,009812 |  |  |
| т | 0,006829 | 0,027046 |  |  |
| пр | 0,006655 | 0,008723 |  |  |
| т | 0,006485 | 0,021962 |  |  |
| ть | 0,006391 | 0,046056 |  |  |
| ли | 0,0063 | 0,035633 |  |  |
| у | 0,006156 | 0,038931 |  |  |
| ет | 0,005995 | 0,04394 |  |  |
| ер | 0,005942 | 0,011792 |  |  |
| го | 0,005923 | 0,01117 |  |  |

Таблиця 4. Біграми з перетином без пробілів.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Біграма** | **Частота** | **Ентропія** | **Загальна ентропія** | **Надлишковість** |
| то | 0,015272 | 0,063233 | 4,173005 | 0,157683 |
| на | 0,013507 | 0,047324 |  |  |
| по | 0,011749 | 0,03973 |  |  |
| ст | 0,011422 | 0,00815 |  |  |
| но | 0,011396 | 0,092135 |  |  |
| не | 0,011307 | 0,04525 |  |  |
| ен | 0,011298 | 0,060234 |  |  |
| ни | 0,009409 | 0,060358 |  |  |
| от | 0,009389 | 0,060267 |  |  |
| ов | 0,009194 | 0,00701 |  |  |
| ко | 0,009174 | 0,008923 |  |  |
| он | 0,009021 | 0,048472 |  |  |
| ал | 0,00885 | 0,004515 |  |  |
| ла | 0,008833 | 0,047155 |  |  |
| ра | 0,008827 | 0,038591 |  |  |
| ос | 0,00868 | 0,026047 |  |  |
| ка | 0,0085 | 0,073565 |  |  |
| ро | 0,008412 | 0,021725 |  |  |
| ет | 0,008012 | 0,034527 |  |  |
| пр | 0,00795 | 0,047341 |  |  |
| ом | 0,007762 | 0,073695 |  |  |
| ли | 0,007746 | 0,02582 |  |  |
| ть | 0,00763 | 0,050662 |  |  |
| ер | 0,007624 | 0,029657 |  |  |
| го | 0,007132 | 0,05545 |  |  |
| ас | 0,007117 | 0,035227 |  |  |
| ва | 0,007097 | 0,006087 |  |  |
| во | 0,006978 | 0,011775 |  |  |
| ол | 0,006765 | 0,031294 |  |  |
| ан | 0,006749 | 0,010979 |  |  |
| ре | 0,006715 | 0,003325 |  |  |
| ес | 0,00652 | 0,054318 |  |  |
| та | 0,006517 | 0,045932 |  |  |
| ат | 0,006488 | 0,045897 |  |  |
| ин | 0,006484 | 0,053637 |  |  |
| од | 0,006373 | 0,014915 |  |  |
| ит | 0,006279 | 0,013729 |  |  |
| те | 0,006273 | 0,004377 |  |  |
| ак | 0,006208 | 0,03898 |  |  |
| ил | 0,006191 | 0,039577 |  |  |

Таблиця 5. Біграми без перетину з пробілами

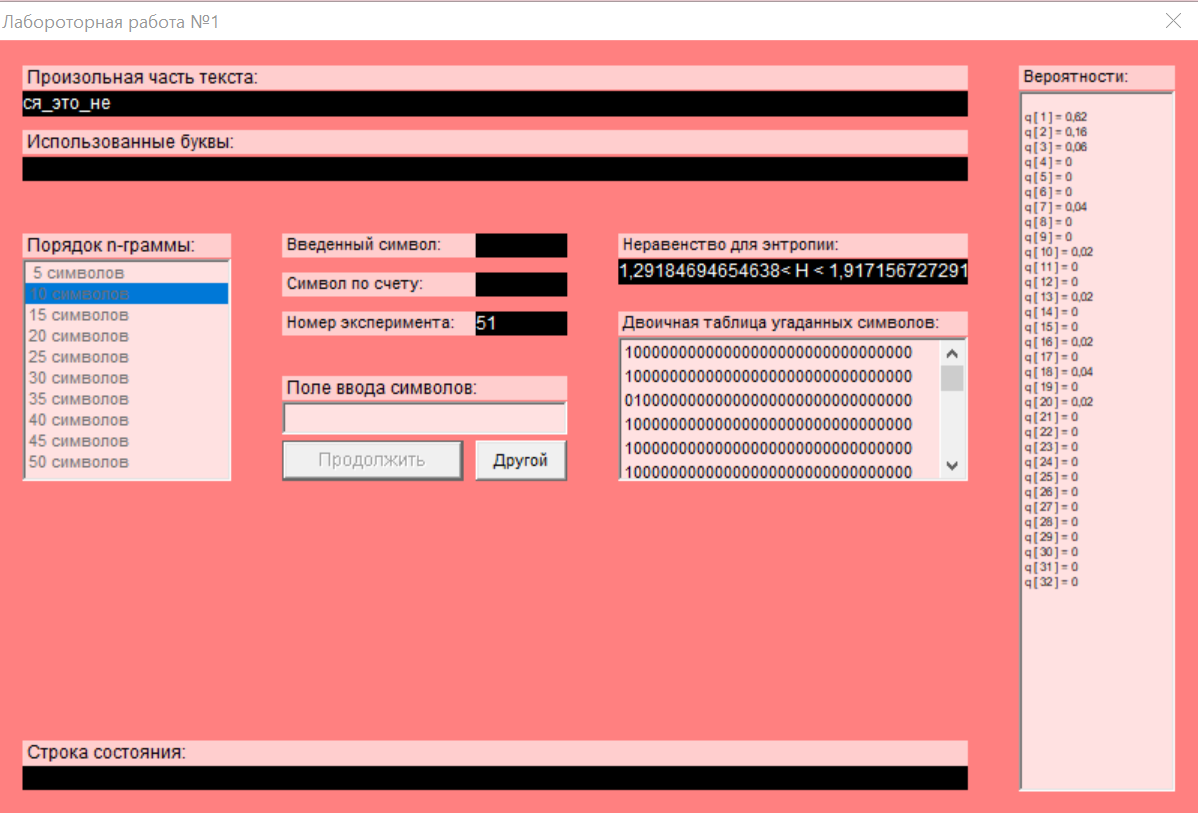
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Біграма** | **Частота** | **Ентропія** | **Загальна ентропія** | **Надлишковість** |
| а | 0,018893 | 0,066298 | 3,993156 | 0,201369 |
| о | 0,018735 | 0,047295 |  |  |
| е | 0,017545 | 0,023039 |  |  |
| п | 0,017126 | 0,077707 |  |  |
| и | 0,016898 | 0,052295 |  |  |
| н | 0,016378 | 0,021754 |  |  |
| я | 0,015185 | 0,001222 |  |  |
| в | 0,01469 | 0,041908 |  |  |
| с | 0,014414 | 0,036553 |  |  |
| то | 0,012231 | 0,108181 |  |  |
| на | 0,011452 | 0,018097 |  |  |
| ь | 0,0108 | 0,016795 |  |  |
| о | 0,009973 | 0,030762 |  |  |
| по | 0,00975 | 0,017524 |  |  |
| не | 0,009496 | 0,047968 |  |  |
| но | 0,00943 | 0,028191 |  |  |
| ст | 0,009318 | 0,009615 |  |  |
| м | 0,009163 | 0,008407 |  |  |
| к | 0,008438 | 0,036024 |  |  |
| и | 0,008207 | 0,044356 |  |  |
| д | 0,00807 | 0,006763 |  |  |
| ни | 0,007821 | 0,027176 |  |  |
| ен | 0,007628 | 0,022641 |  |  |
| ла | 0,007473 | 0,027352 |  |  |
| ра | 0,007385 | 0,009245 |  |  |
| й | 0,007324 | 0,037411 |  |  |
| ко | 0,00723 | 0,000727 |  |  |
| ал | 0,007141 | 0,10049 |  |  |
| от | 0,006994 | 0,049955 |  |  |
| ро | 0,006973 | 0,019008 |  |  |
| т | 0,006935 | 0,034189 |  |  |
| ка | 0,006895 | 0,038201 |  |  |
| м | 0,006732 | 0,054735 |  |  |
| пр | 0,006628 | 0,091736 |  |  |
| т | 0,006512 | 0,013361 |  |  |
| ть | 0,006324 | 0,007928 |  |  |
| ли | 0,006273 | 0,045897 |  |  |
| у | 0,006067 | 0,051418 |  |  |
| ер | 0,006012 | 0,044053 |  |  |
| го | 0,005981 | 0,089448 |  |  |

Таблиця 6. Біграми без перетину без пробілів

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Біграма** | **Частота** | **Ентропія** | **Загальна ентропія** | **Надлишковість** |
| то | 0,015325 | 0,062592 | 4,172907 | 0,157703 |
| на | 0,013355 | 0,039521 |  |  |
| по | 0,011589 | 0,092379 |  |  |
| но | 0,011513 | 0,060358 |  |  |
| ен | 0,011456 | 0,060164 |  |  |
| ст | 0,011389 | 0,008819 |  |  |
| не | 0,011289 | 0,004397 |  |  |
| ни | 0,009501 | 0,047289 |  |  |
| от | 0,009268 | 0,025496 |  |  |
| ко | 0,009174 | 0,021061 |  |  |
| ов | 0,009156 | 0,047289 |  |  |
| ра | 0,00885 | 0,026082 |  |  |
| ал | 0,008838 | 0,029384 |  |  |
| ла | 0,008814 | 0,048379 |  |  |
| он | 0,008805 | 0,07353 |  |  |
| ос | 0,008662 | 0,005739 |  |  |
| ка | 0,008598 | 0,031582 |  |  |
| ро | 0,008405 | 0,003294 |  |  |
| пр | 0,007941 | 0,045155 |  |  |
| ет | 0,007808 | 0,053698 |  |  |
| ом | 0,007793 | 0,047201 |  |  |
| ли | 0,00772 | 0,004838 |  |  |
| ер | 0,007635 | 0,039632 |  |  |
| ть | 0,007526 | 0,005042 |  |  |
| го | 0,007117 | 0,007472 |  |  |
| ас | 0,007111 | 0,0554 |  |  |
| во | 0,007056 | 0,017698 |  |  |
| ва | 0,00705 | 0,011212 |  |  |
| ан | 0,006935 | 0,033212 |  |  |
| ол | 0,006735 | 0,073865 |  |  |
| ре | 0,006699 | 0,020201 |  |  |
| та | 0,006511 | 0,030966 |  |  |
| ес | 0,006511 | 0,010912 |  |  |
| ат | 0,006496 | 0,054173 |  |  |
| од | 0,006472 | 0,062092 |  |  |
| те | 0,006466 | 0,050395 |  |  |
| ин | 0,00635 | 0,028548 |  |  |
| ор | 0,006266 | 0,027447 |  |  |
| ит | 0,006147 | 0,049737 |  |  |
| ак | 0,006111 | 0,011643 |  |  |

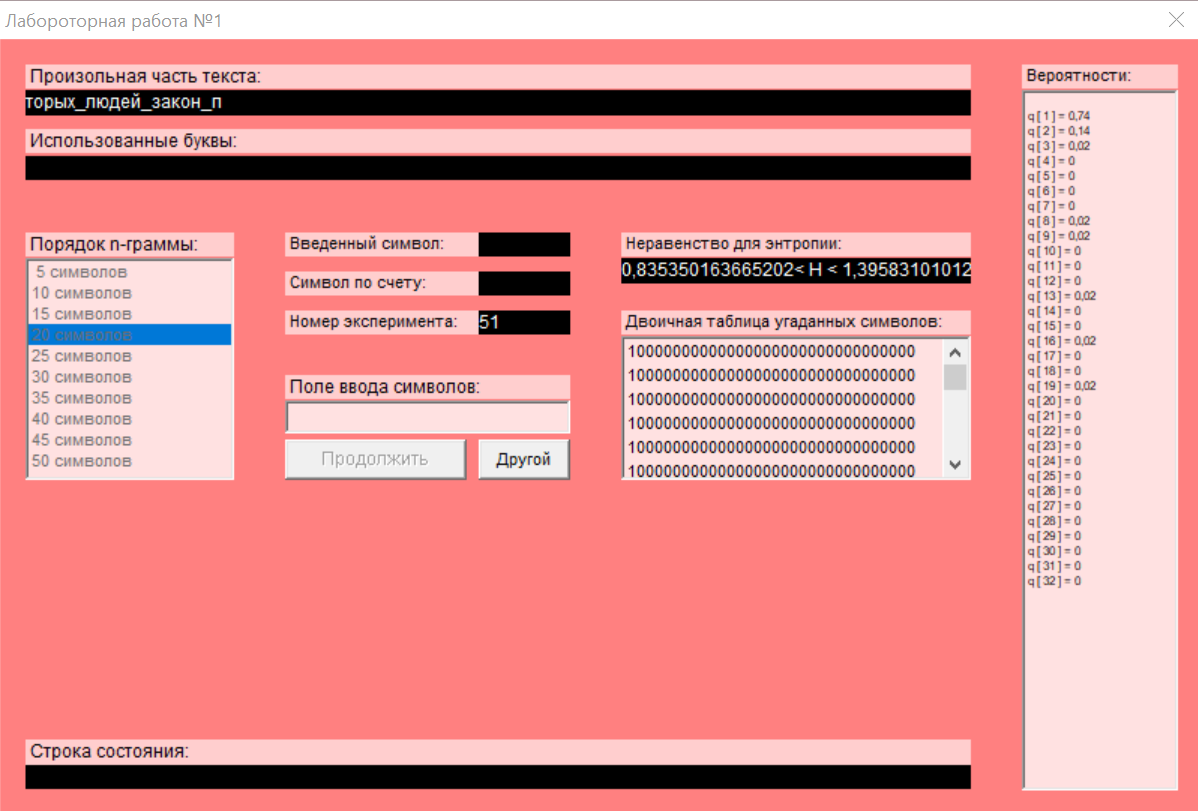
Далі за допомогою CoolPinkProgram я оцінила значення H10, H20, H30:

H10



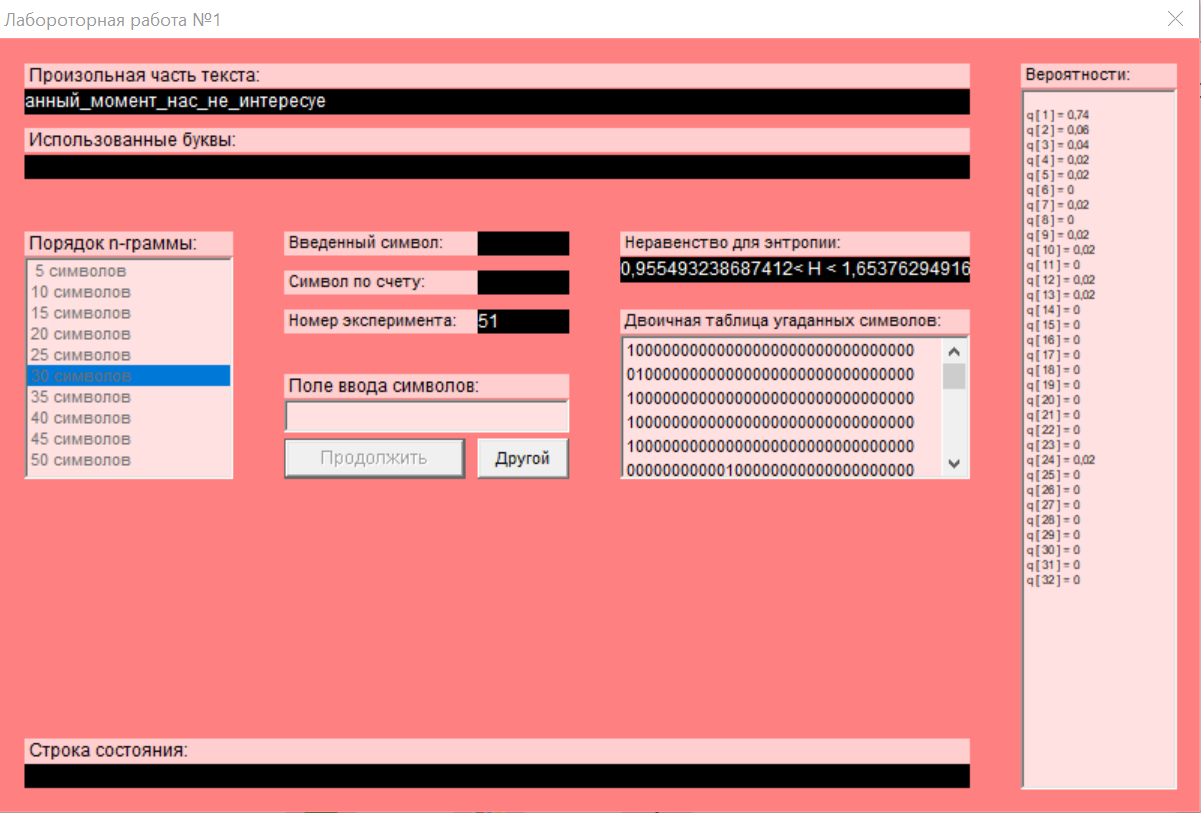
1,29185 < H10 < 1,91716

H20



0,83535 < H20 < 1,39583

H30



0,95549 < H30 < 1,65376

Тепер визначу надлишковість.

= = 5

Отже для Н10:

1) R = 1 - = 0,74163

2) R = 1 - = 0,616568

H20:

1) R = 1 - = 0,83293

2) R = 1 - = 0,720834

H30:

1) R = 1 - = 0,808902

2) R = 1 - = 0,669248

**Висновки**

Після виконання роботи я краще зрозуміла поняття ентропії та надлишковості. Під час виконання практикуму я змогла обчислити частоти монограм та біграм у великому тексті, звідси змогла визначити ентропію та надлишковість. А за допомогою CoolPinkProgram оцінила значення H10, H20 та H30.