Міністерство освіти і науки України
Національний технічний університет України
“Київський політехнічний інститут ім. Ігоря Сікорського”
Фізико-технічний інститут

**Лабораторна робота № 1**

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

**Виконали:**
Студентки 3 курсу,
Групи ФБ-06 Вєрнікова
Лілія
Товкач Катерина

**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**

0. Уважно прочитати методичні вказівки до виконання комп'ютерного практикуму.

1. Написати програми для підрахунку частот букв і частот біграм в тексті, а також

підрахунку 1H та 2H за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення 1H та 2H на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення 1H та 2H 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**

Знайшли текст, який будемо досліджувати, потім зчитали його та очистили від зайвих смволів, отриманий текст зберегли у файлі, також зберегли цей самий текст, але без пробілів у іншому файлі. Для очистки використали фільтр де перевіряли чи належить символ з тексту до нашого алфавіту. Після ми спочатку написали функції для пошуку монограм, але в подальшому вирішили оптимізувати код і зробили одну велику функцію де містяться всі інші функції що ми будемо використовувати. У цій функції ми одразу шукали монограму і біграми. Отримані дані записуємо в csv використовуючи бібліотеку pandas.

## Результати

**Текст з пробілами: монограми:**

Енитропія: 4.374552813221424

Надлишковість: 0.1250894373557152

**біграми з перетином:**

Енитропія: 3.982927743685439

Надлишковість: 0.20341445126291213

**біграми без перетину:**

Енитропія: 3.9827166848356628

Надлишковість: 0.20345666303286747

**Текст без пробілів:**

**монограми:**

Енитропія: 4.449295156504394

Надлишковість:

0.10191383672542709 **біграми з**

**перетином:** Енитропія:

4.143456816967816

Надлишковість: 0.16364702620267146

**біграми без перетину:**

Енитропія: 4.143254433060141

Надлишковість: 0.16368787720957445 Найчастіша поява монограм

Монограми без пробілів

| Буква | к-сть | Частота |
|---|---|---|
| о | 51247 | 0.11113183283926469 |
| а | 40412 | 0.08763556166605586 |
| е | 37479 | 0.08127519587454488 |
| и | 30990 | 0.06720345580597523 |

| | | |
|---|---|---|
| **н** | 29234 | 0.0633954768322646 |
| **т** | 27982 | 0.060680448543491416 |
| **л** | 24261 | 0.05261126303029252 |
| **с** | 23145 | 0.050191157942216735 |
| **р** | 22203 | 0.04814838106679794 |
| **в** | 21527 | 0.04668243927509612 |
| **к** | 16813 | 0.036459880686216896 |
| **м** | 13893 | 0.03012770608300787 |
| **у** | 13698 | 0.0297048382584785 |
| **д** | 13103 | 0.028414549255427347 |
| **п** | 12855 | 0.02787674812474384 |
| **г** | 9139 | 0.019818405376276466 |
| **я** | 8841 | 0.019172176598277735 |
| **ь** | 8640 | 0.018736297456070538 |
| **з** | 8298 | 0.017994652348434413 |
| **ы** | 8030 | 0.017413480158824816 |
| **ч** | 7150 | 0.015505153566076893 |
| **б** | 7104 | 0.015405400130546887 |
| **й** | 5220 | 0.011319846379709285 |
| **ж** | 4160 | 0.009021180256626555 |
| **ш** | 3973 | 0.00861566085566762 |
| **х** | 3760 | 0.008153759078104771 |
| **ю** | 2531 | 0.0054886750709659 |
| **щ** | 1664 | 0.003608472102650622 |
| **ц** | 1474 | 0.003196447042852775 |
| **э** | 1414 | 0.0030663338660745073 |
| **ф** | 897 | 0.001945191992835101 |

# Монограми з пробілами

| Буква | к-сть | Частота |
|---|---|---|
| | 86933 | 0.1586166000693342 |
| **о** | 51247 | 0.09350447935482693 |
| **а** | 40412 | 0.0737351068294196 |
| **е** | 37479 | 0.06838360063495538 |
| **и** | 30990 | 0.05654387213312168 |

| | | |
|---|---|---|
| **н** | 29234 | 0.05333990183735654 |
| **т** | 27982 | 0.05105552210484062 |
| **л** | 24261 | 0.04426624336307406 |
| **с** | 23145 | 0.04223000711587936 |
| **р** | 22203 | 0.04051124856313974 |
| **в** | 21527 | 0.03927782947433722 |
| **к** | 16813 | 0.030676738372835587 |
| **м** | 13893 | 0.02534895177623296 |
| **у** | 13698 | 0.024993157808309158 |
| **д** | 13103 | 0.02390753006002883 |
| **п** | 12855 | 0.023455033116207783 |
| **г** | 9139 | 0.016674877296695676 |
| **я** | 8841 | 0.016131151130330066 |
| **ь** | 8640 | 0.015764409655700914 |
| **з** | 8298 | 0.015140401773496086 |
| **ы** | 8030 | 0.014651413140657215 |
| **ч** | 7150 | 0.013045778823872863 |
| **б** | 7104 | 0.012961847939131863 |
| **й** | 5220 | 0.009524330833652636 |
| **ж** | 4160 | 0.007590271315707848 |
| **ш** | 3973 | 0.007249074023391173 |
| **х** | 3760 | 0.006860437535351324 |
| **ю** | 2531 | 0.004618023245205904 |
| **щ** | 1664 | 0.003036108526283139 |
| **ц** | 1474 | 0.0026894374806137903 |
| **э** | 1414 | 0.0025799624135603117 |
| **ф** | 897 | 0.0016366522524495046 |

## Біграми

## Неперехресна та перехресна біграми з пробілами

| Буква | к-сть | Частота |
|---|---|---|
| о | 5773 | 0.02106665206999106 |
| а | 4911 | 0.017921068476654443 |
| п | 4574 | 0.0166912985567537 |
| в | 4512 | 0.016465050084843176 |
| и | 4424 | 0.016143923221486306 |
| е | 4384 | 0.015997956465415002 |
| н | 4175 | 0.015235280164942435 |
| с | 4048 | 0.014771835714416042 |
| то | 3564 | 0.013005637965953254 |
| и | 2955 | 0.01078329410476764 |
| к | 2694 | 0.009830861021402376 |
| о | 2668 | 0.009735982629956027 |
| но | 2633 | 0.009608261718393636 |
| ь | 2594 | 0.009465944131224114 |
| на | 2565 | 0.009360118233072418 |
| ст | 2480 | 0.009049938876420895 |
| по | 2460 | 0.008976955498385243 |
| я | 2459 | 0.00897330632948346 |
| ко | 2377 | 0.008674074479537285 |
| не | 2366 | 0.008633933621617677 |
| ал | 2315 | 0.008447826007626762 |
| ро | 2212 | 0.008071961610743153 |
| л | 2131 | 0.007776378929698761 |
| ла | 2122 | 0.007743536409582717 |

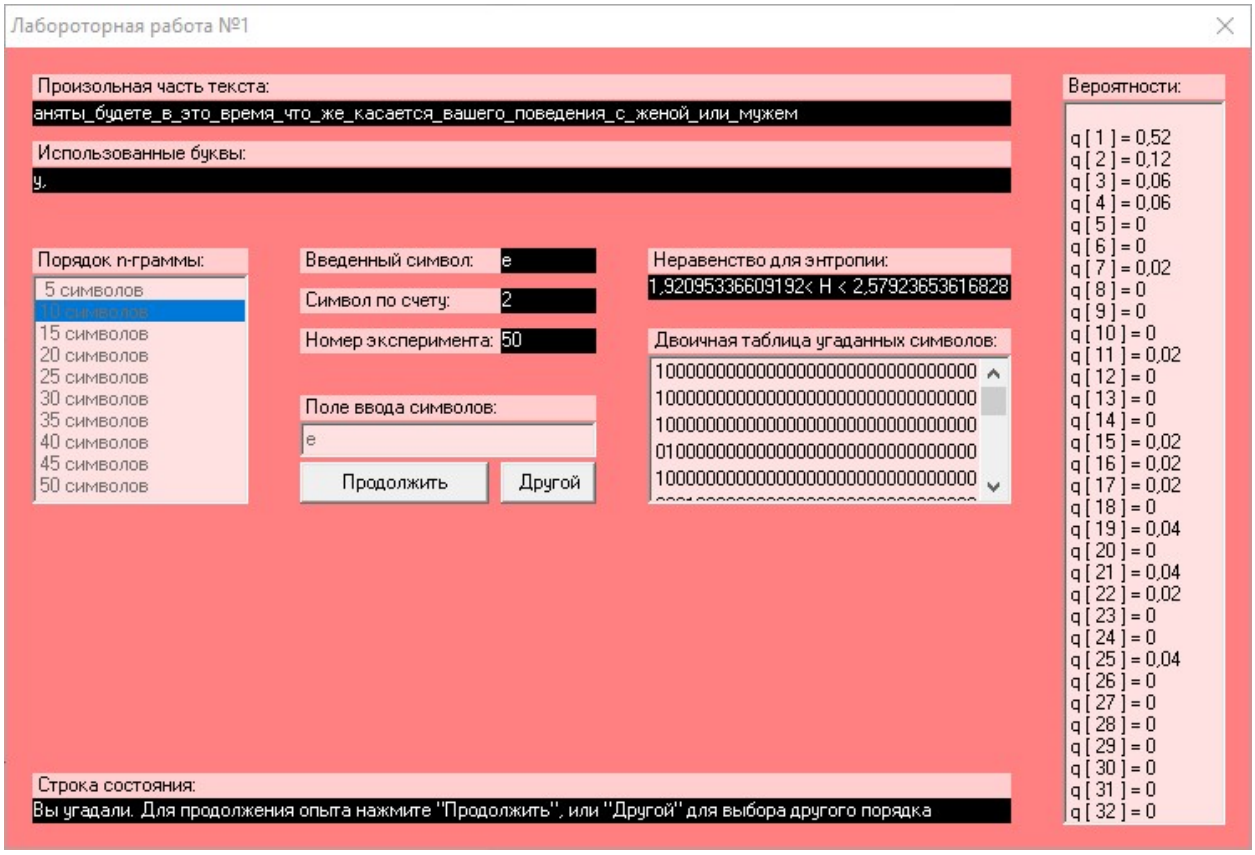
| Буква | к-сть | Частота |
|---|---|---|
| о | 11312 | 0.020639736967425635 |
| а | 9955 | 0.01816377134995776 |
| п | 9151 | 0.016696802774833097 |
| в | 8866 | 0.01617679525753144 |
| и | 8857 | 0.01616037396751139 |
| е | 8833 | 0.01611658386079125 |
| н | 8382 | 0.015293694772008634 |
| с | 7993 | 0.014583930125586377 |
| то | 7214 | 0.013162576244961857 |
| и | 5890 | 0.010746822024234175 |
| к | 5384 | 0.009823580607551239 |
| о | 5319 | 0.009704982401850863 |
| но | 5306 | 0.009681262760710787 |
| ь | 5214 | 0.009513400684950253 |
| я | 5098 | 0.00930174850246958 |
| ст | 5073 | 0.009256133807969434 |
| на | 5049 | 0.009212343701249296 |
| по | 4944 | 0.009020761984348686 |
| не | 4783 | 0.008727003351767753 |
| ко | 4760 | 0.008685037832827619 |
| ал | 4713 | 0.008599282207167346 |
| ро | 4444 | 0.008108468094345785 |
| л | 4268 | 0.007787340645064764 |
| го | 4250 | 0.007754498065024659 6 |

**Неперехресна та перехресна біграми без пробілів**

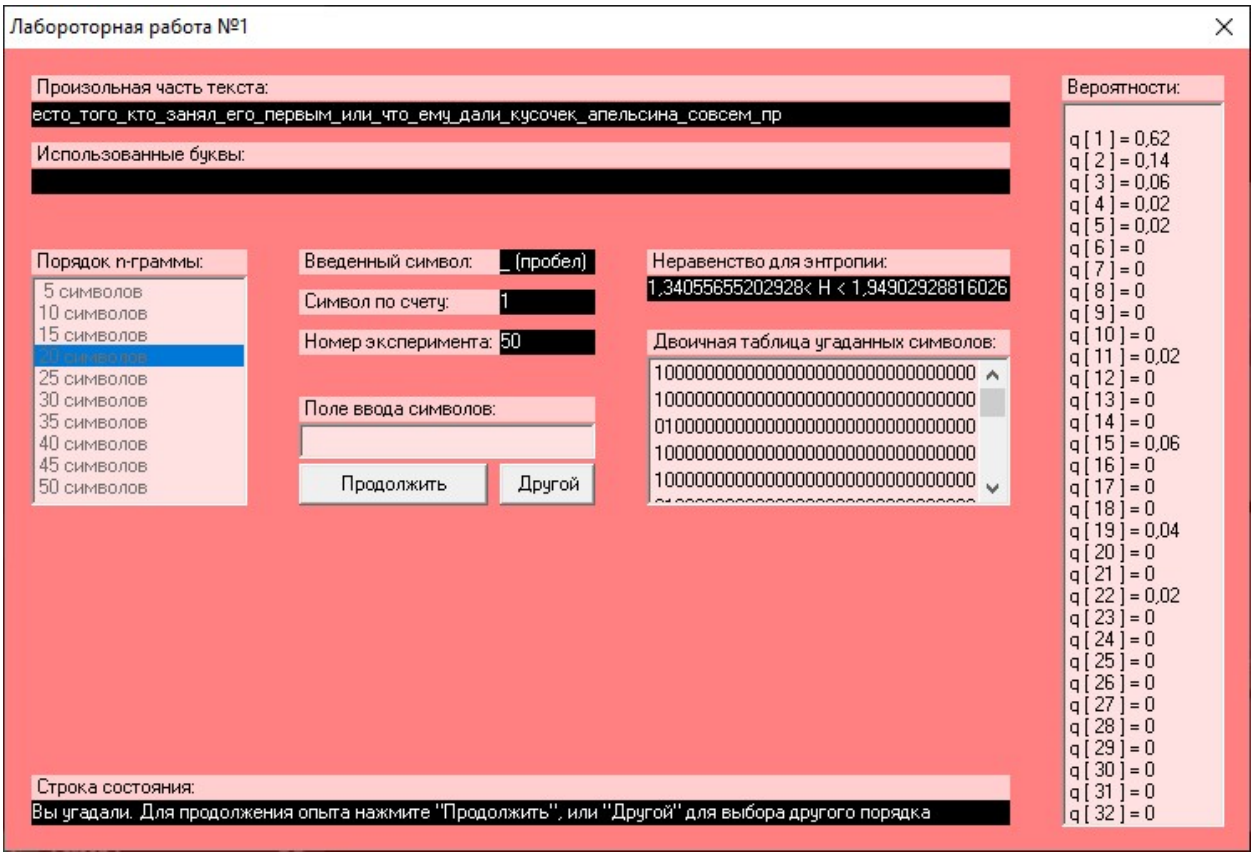
| Буква | к-сть | Частота |
|---|---|---|
| то | 3691 | 0.016008292564449533 |
| но | 2698 | 0.011701537073661566 |
| ст | 2591 | 0.011237465736789147 |
| на | 2528 | 0.010964227473023143 |
| по | 2503 | 0.010855799590576316 |
| ов | 2494 | 0.010816765552895458 |
| ко | 2478 | 0.01074737170812949 |
| ал | 2383 | 0.010335345754831547 |
| не | 2369 | 0.010274626140661323 |
| ро | 2305 | 0.009997050761597447 |
| ла | 2254 | 0.00977585788140592 |
| го | 2193 | 0.009511293848235661 |
| от | 2127 | 0.009225044238576039 |
| ра | 2113 | 0.009164324624405815 |
| ос | 2079 | 0.009016862704278131 |
| ен | 2037 | 0.00883470386176746 |
| он | 2021 | 0.008765310017001492 |
| ка | 1991 | 0.0086351965580653 |
| ни | 1960 | 0.008500745983831234 |
| ло | 1909 | 0.008279553103639706 |
| во | 1891 | 0.008201485028277992 |
| ор | 1874 | 0.008127754068214149 |
| ан | 1841 | 0.007984629263384339 |
| ол | 1829 | 0.00793258387980986 |

| Буква | к-сть | Частота | Ентропія |
|---|---|---|---|
| то | 7439 | 0.016131900350438915 | 0.09604836339559475 |
| но | 5415 | 0.011742739668991361 | 0.07529547014335537 |
| ст | 5165 | 0.011200600256757226 | 0.07258302739160925 |
| на | 5076 | 0.011007598626001874 | 0.07160835238127837 |
| ов | 5027 | 0.010901339301203983 | 0.07106965508688676 |
| по | 4946 | 0.010725686131640124 | 0.07017587172909275 |
| ал | 4907 | 0.010641112383331599 | 0.0697440555791034 |
| ко | 4887 | 0.010597741230352868 | 0.06952223571450408 |
| не | 4803 | 0.010415582387842199 | 0.06858778364937902 |
| ро | 4521 | 0.009804049130842095 | 0.06541660183999268 |
| ла | 4378 | 0.00949394538704417 | 0.06378769747991177 |
| го | 4264 | 0.009246729815065404 | 0.06247868359499733 |
| от | 4259 | 0.009235887026820721 | 0.062421054260072696 |
| ра | 4168 | 0.009038548280767496 | 0.0613689698967814 |
| ен | 4109 | 0.00891060337948024 | 0.060683536088198736 |
| ос | 4093 | 0.008875906457097255 | 0.06049720047809504 |
| он | 4090 | 0.008869400784150445 | 0.060462240789665715 |
| ни | 4028 | 0.008734950209916381 | 0.059738191758063795 |
| ка | 3907 | 0.008472554734395059 | 0.05831648606624988 |
| ор | 3798 | 0.008236181950660976 | 0.05702574727997404 |
| во | 3775 | 0.008186305124735435 | 0.05675214841338552 |
| ан | 3723 | 0.008073540126990736 | 0.05613195680458713 |
| ол | 3722 | 0.0080713715693418 | 0.056120007874980705 |
| ло | 3709 | 0.008043180319905625 | 0.0559645952167137 |

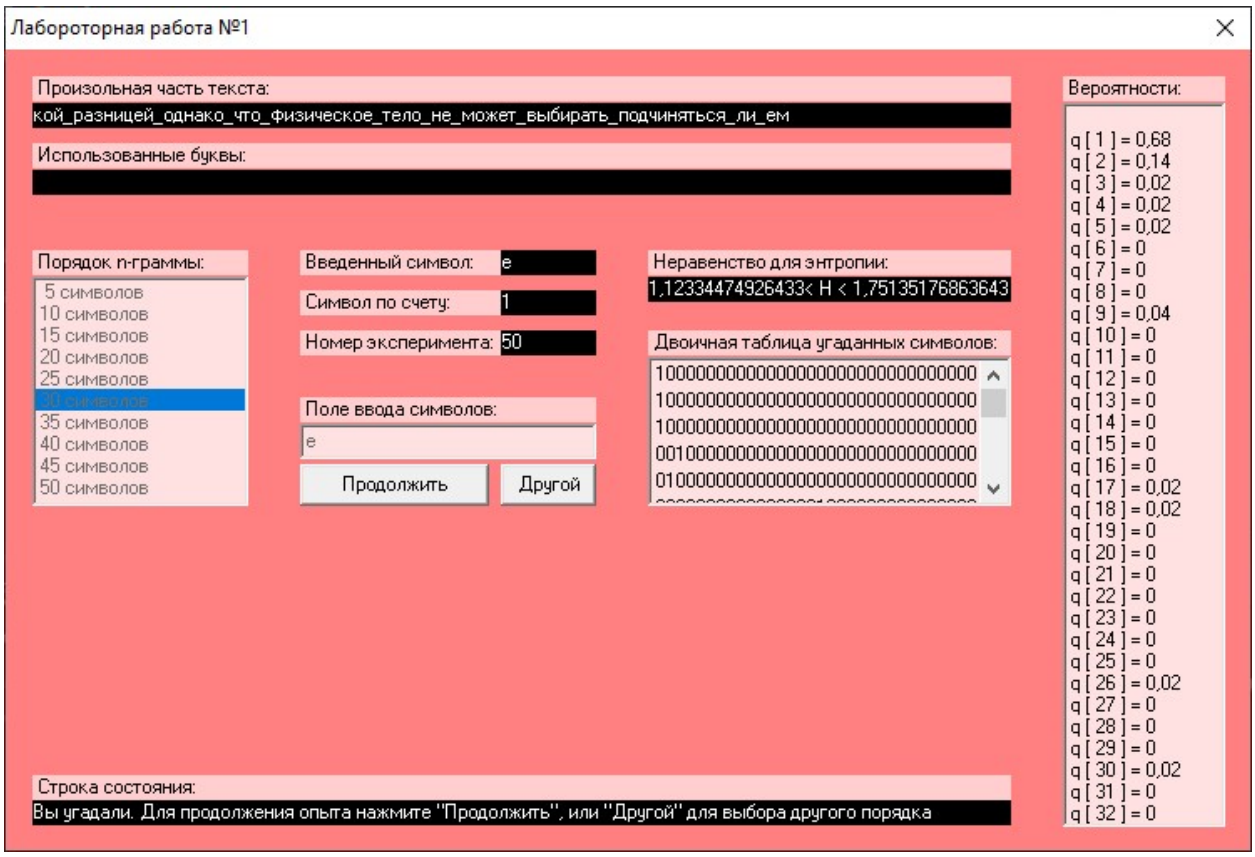
Cool Pink




R= 0.54998100977398




R = 0.671041415981046




R = 0.712530348209924

Висновки: засвоїли поняття ентропії на символ джерела та його надлишковості, вивчили та порівняли різні моделі джерела відкритого тексту для наближеного визначення ентропії, набули практичних навичок щодо оцінки ентропії на символ джерела. Після аналізу тексту визначили що найчастіші у використанні букви рос. алфавіту «о», «а», «е».