Міністерство освіти і науки України Національний технічний університет   
України "Київський політехнічний інститут імені Ігоря Сікорського"   
Фізико-технічний інститут

**Криптографія**

**Комп’ютерний практикум №1**

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

**Виконали:**

Студенти ІІІ курсу

групи ФБ-95

Пашинський М. О.

Бурчак Б. Ю.

**Перевірила:**

Селюх К. І.

Київ – 2021

**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.

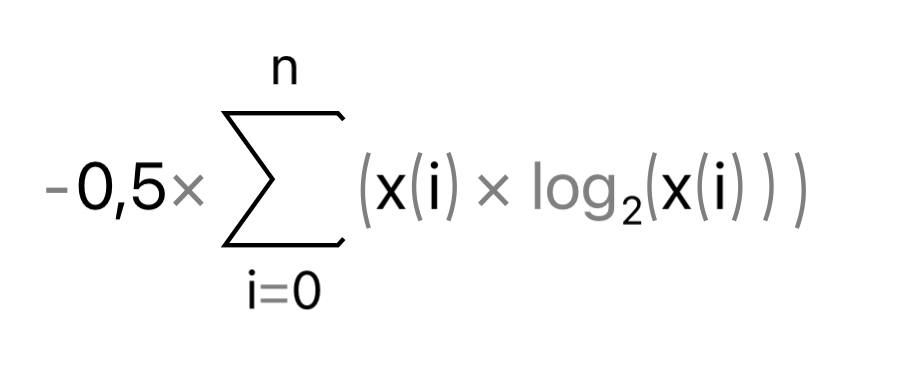
**Завдання:**

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:**

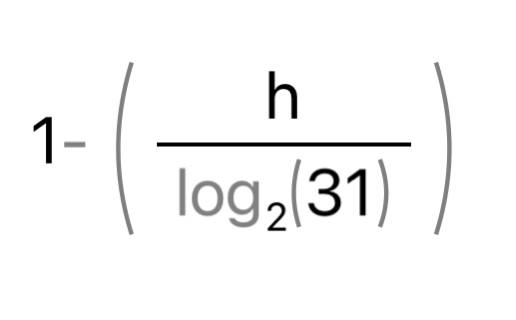
За основу тексту було взято науково-фантастичний роман «Дюна». Програма була розроблена на мові Python, вона фільтрує текст від усього лишнього і підраховує частоти монограм та біграм. Інформація записується в окремі файли. За допомогою отриманих даних було підраховано значення ентропій і надлишковості за даними формулами:

Ентропія(**H**):



----------------------------------------------------------------------------------------------------------------

Надлишковість(**R**):



**Монограми і їх частоти:**

|  |  |
| --- | --- |
| Символи з пробілом | Символи без пробілу |
| а --> 0.06932653754842562  б --> 0.014045151669829198  в --> 0.036196250027695664  г --> 0.015252473693297566  д --> 0.02611210118763197  е --> 0.07135448738770714  ж --> 0.009208338715698657  з --> 0.014073578919689257  и --> 0.058503280212000394  й --> 0.00925516006840934  к --> 0.02652680459735518  л --> 0.043232412415852205  м --> 0.026674793515744307  н --> 0.05660994176176209  о --> 0.09319163185373679  п --> 0.02546663539669182  р --> 0.0376268095721239  с --> 0.04626911157737375  т --> 0.05319240106167417  у --> 0.023863004066350873  ф --> 0.0017980235536487002  х --> 0.007820001998268445  ц --> 0.003035863065937424  ч --> 0.011567800454083512  ш --> 0.006299144130755316  щ --> 0.0034004007406134686  ы --> 0.016297593173446773  ь --> 0.01554803348228376  э --> 0.0032310913848292966  ю --> 0.005216818397112795  я --> 0.017345638988140403  \_ --> 0.15245868538183022 | а --> 0.08179723672781458  б --> 0.016571642499996302  в --> 0.04270735762775485  г --> 0.01799614181660163  д --> 0.030809236950762535  е --> 0.08418998125165546  ж --> 0.010864766775230482  з --> 0.01660518334263105  и --> 0.06902705414231873  й --> 0.01092001051604066  к --> 0.03129853865508127  л --> 0.05100920942754167  м --> 0.031473148335856294  н --> 0.06679313537330711  о --> 0.10995526736737433  п --> 0.030047662523879352  р --> 0.04439525120858014  с --> 0.05459216061723043  т --> 0.06276083554184984  у --> 0.02815556440113072  ф --> 0.0021214582966480365  х --> 0.009226691210671512  ц --> 0.0035819646943171454  ч --> 0.013648656713914863  ш --> 0.007432256129712303  щ --> 0.0040120766763392535  ы --> 0.019229261031114556  ь --> 0.018344867930465886  э --> 0.00381231136358816  ю --> 0.006155237871162718  я --> 0.020465832979428118 |

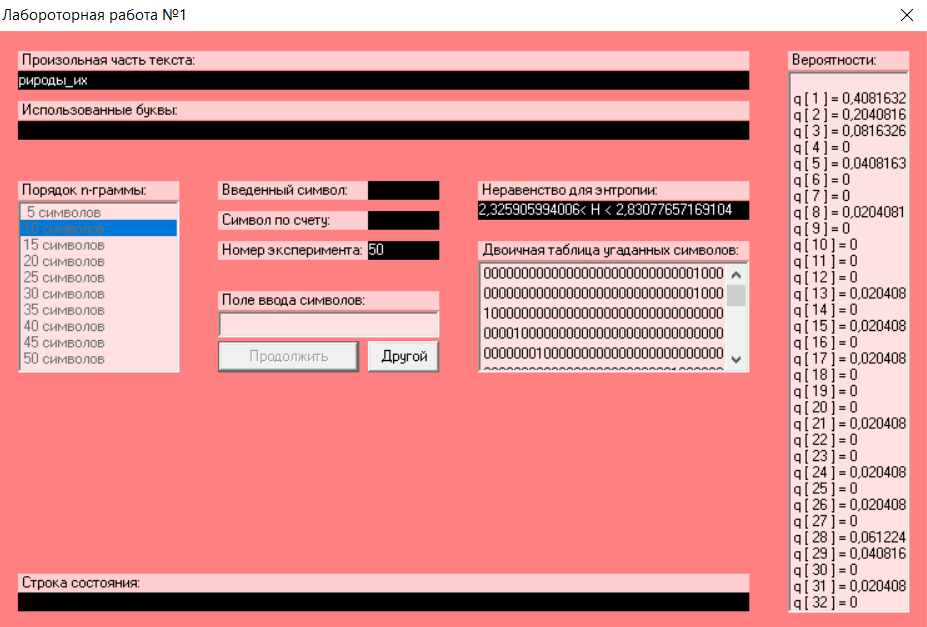
**Перші 10 бірам і їх частот (за найбільшою частотою появи):**

|  |  |  |  |
| --- | --- | --- | --- |
| Частоти з кроком 1 1 1 без пробілів | Частоти з кроком 2 без пробілів | Частоти з кроком 1 з пробілами | Частоти з кроком 2 з пробілами |
| то : 0.015386  ст : 0,013524  на : 0,012524  по : 0,011953  но : 0,011815  ен : 0,011453  ни : 0,010609  ов : 0,010382  ал : 0,010292  не : 0,010067 | то : 0,015386  ст : 0,013524  на : 0,012524  по : 0,011954  но : 0,011815  ен : 0,011453  ни : 0,010609  ов : 0,010382  ал : 0,010292  не : 0,010067 | о\_ : 0,019159  п\_ : 0,017552  а\_ : 0,016985  е\_ : 0,015536  и\_ : 0,015524  c\_ : 0,015309  в\_ : 0,013683  н\_ : 0,01337  то : 0,012565  о\_ : 0,01178 | о\_ : 0,019159  п\_ : 0,017552  а\_ : 0,016985  е\_ : 0,015536  и\_ : 0,015524  с\_ : 0,015309  в\_ : 0,013683  н\_ : : 0,01337  то : 0,012565  о\_ : 0,01178 |

|  |  |  |
| --- | --- | --- |
|  | **Ентропія:** | **Надлишковість:** |
| Монограми без пробілу. | 2,229394 | 0,549999 |
| Монограми з пробілом. | 2,197483 | 0,560503 |
| Біграми без перетину та без пробілів. | 4,14984 | 0,162359 |
| Біграми без перетину з пробілами. | 4,003918 | 0,199216 |
| Біграми з перетином та без пробілів. | 4,149933 | 0,16234 |
| Біграми з перетином та пробілами. | 4,00399 | 0,199202 |

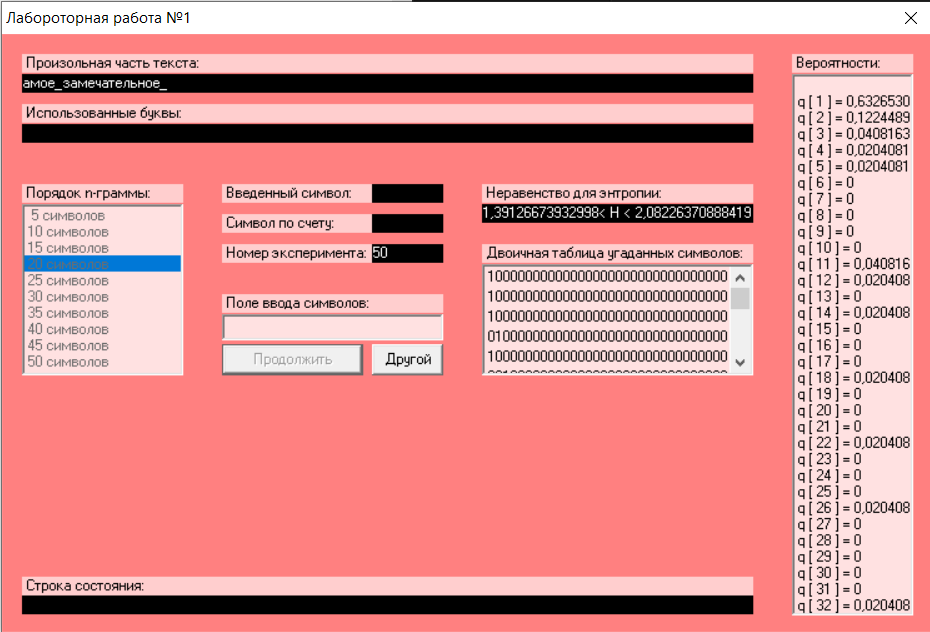
**Ентропія для 10 символів:**

H(10) = 2.57834128284852 (середнє значення)



**Ентропія для 20 символів:**

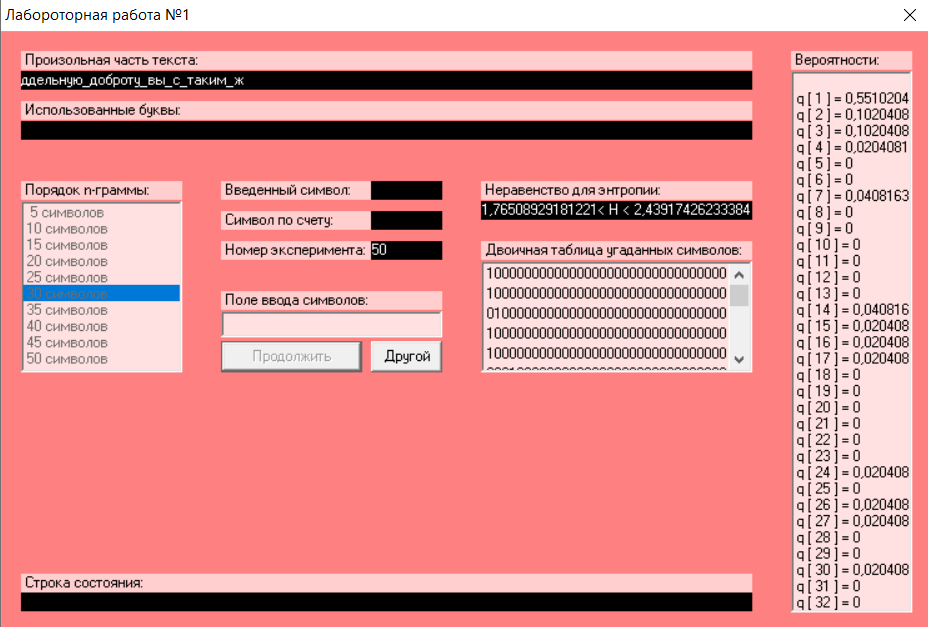
H(20) = 1.736765224107085 (середнє значення)



**Ентропія для 30 символів:**

H(30) = 2.102131777073025 (середнє значення)

|  |  |  |
| --- | --- | --- |
|  | Ентропія: | Надлишковість: |
| H(10) | 2.57834128284852 | 0.484332 |
| H(20) | 1.736765224107085 | 0.652647 |
| H(30) | 2.102131777073025 | 0.579574 |



**Висновок:**

В ході цієї лабораторної роботи ми засвоїли поняття ентропії та надлишковості, навчилися знаходити частоти біграм та монограм в тексті. А також вдосконалили свої практичні навички щодо мови програмування Python.