СИМЕТРИЧНА КРИПТОГРАФІЯ КОМП'ЮТЕРНИЙ ПРАКТИКУМ №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

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.

Постановка задачі:

1. Уважно прочитати методичні вказівки до виконання комп'ютерного практикуму
2. Написати програми для підрахунку частот букв і частот біграм в тексті, а також підрахунку H_1 та H_2 за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення H_1 та H_2 на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Так одержати значення H_1 та H_2 на тому ж тексті, в якому вилучено всі пробіли.
3. За допомогою програми CoolPinkProgram оцінити значення $H^{(10)}$, $H^{(20)}$, $H^{(30)}$.
4. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

Хід роботи:

1. Написати програму для підрахунку частот букв і частот біграм в тексті роману "Мастер и Маргарита", підрахувати також H_1 та H_2 на цьому тексті, враховуючи і не враховуючи пробіли.

!!!УСІ ТАБЛИЦІ ЗНАХОДЯТЬСЯ В ОКРЕМОМУ ФАЙЛІ "charts.pdf"!!!

Таблиця частот монограм, враховуючи пробіли:

$H_1 = 4.379021219510$

Таблиця частот монограм, не враховуючи пробіли:

$H_1 = 4.450124790210359$

Таблиця частот біграм, що перетинаються, враховуючи пробіли:

$H_2 = 7.963835600458476$

Таблиця частот біграм, що перетинаються, не враховуючи пробіли:

$H_2 = 8.293071816295$

Таблиця частот біграм, що не перетинаються, враховуючи пробіли:

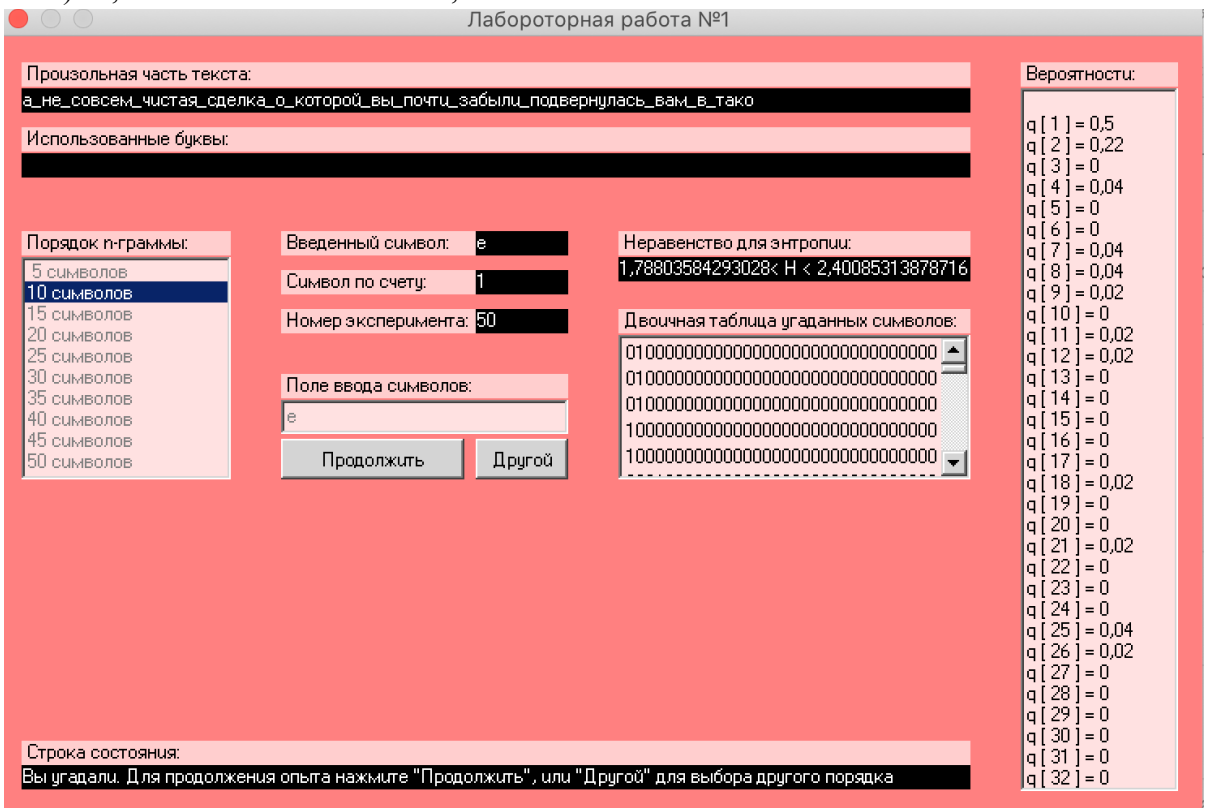
$H_2 = 7.893119625550205$

Таблиця частот біграм, що не перетинаються, не враховуючи пробіл:

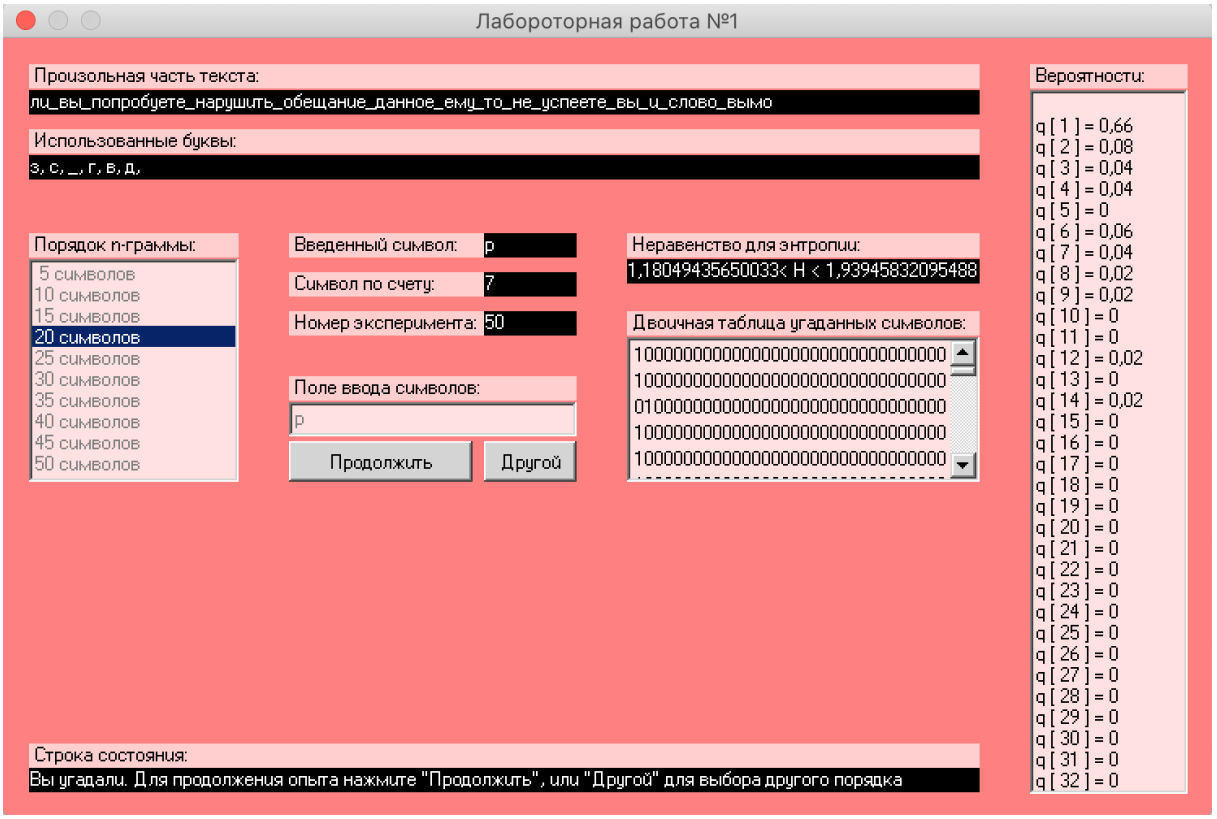
$H_2 = 8.292505499297492$

2. Провести по 50 експериментів за допомогою програми CoolPinkProgram для знаходження значень $H^{(10)}$, $H^{(20)}$, $H^{(30)}$, тобто інтуїтивно вгадувати наступний символ у тексті, якщо відомі відповідно 9, 19 та 29 попередніх.

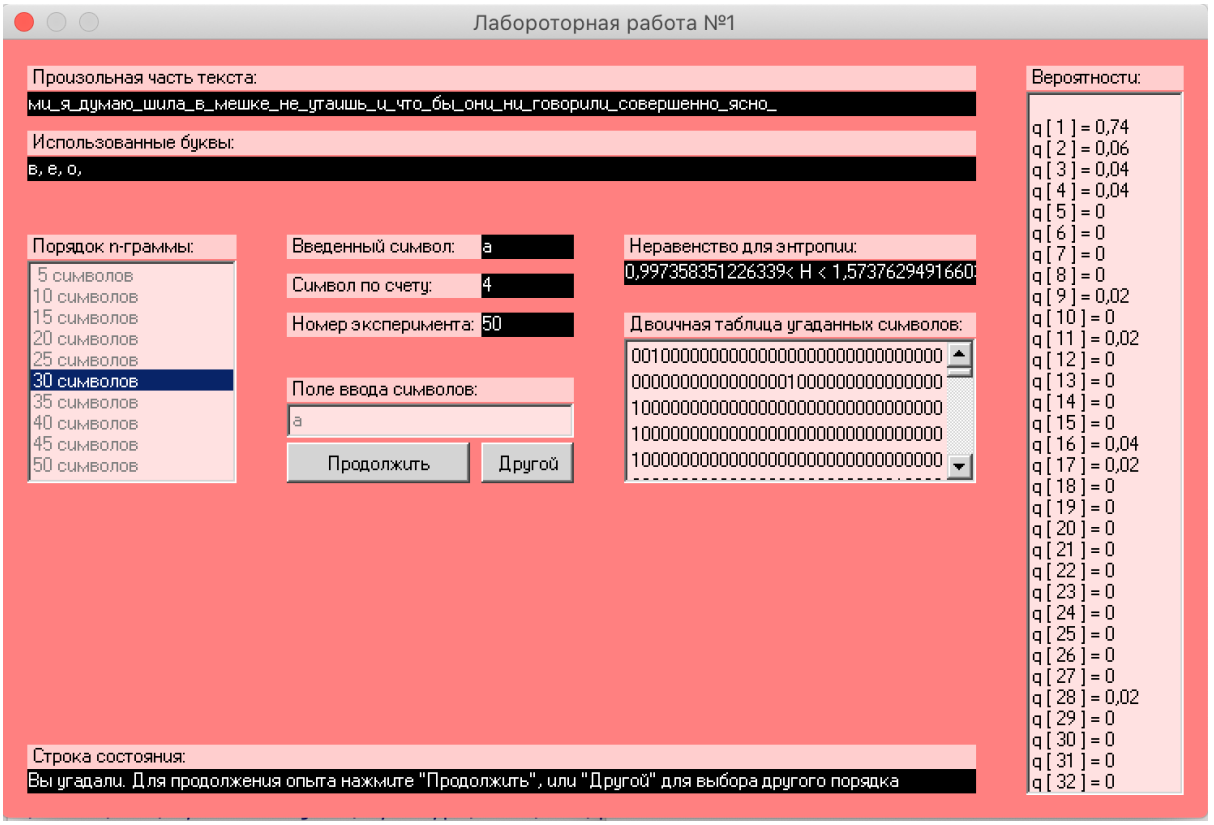
1) 1,78803584293028 < H < 2,40085313878716



1) 1,18049435650033 < H < 1,93945832095488



2) 0,997358351226339 < H < 1,57376294916603



3. Оцінити надлишковість російської мови для кожної з 9-ти моделей джерела.

For monograms $H_0 = 5.0$

For bigrams $H_0 = 10.0$

Monogram entropy considering spaces:

$H_1 = 4.379021219510967$

R = 0.12419575609780664

Crossing bigram entropy considering spaces:

$H_2 = 3.981917800229238$

R = 0.6018082199770762

No crossing bigram entropy considering spaces:

$H_2 = 3.9465598127751025$

R = 0.6053440187224898

Monogram entropy not considering spaces:

$H_1 = 4.450124790210359$

R = 0.10997504195792818

Crossing bigram entropy not considering spaces:

$H_2 = 4.1465359081475$

R = 0.5853464091852499

No crossing bigram entropy not considering spaces:

$H_2 = 4.146252749648746$

R = 0.5853747250351253

For $H^{(10)}$:

$$1.78803584293028 < H^{(10)} < 2.40085313878716$$
$$0.519829372242568 < R < 0.642392831413944$$

For $H^{(20)}$:

$$1.18049435650033 < H^{(20)} < 1.93945832095488$$
$$0.612108335809024 < R < 0.763901128699934$$

For $H^{(30)}$:

$$0.997358351226339 < H^{(30)} < 1.57376294916603$$
$$0.685247410166794 < R < 0.8005283297547322$$

Висновок:

Я засвоїла поняття ентропії на символ джерела та його надлишковості, за допомогою деякого тексту, дізналася частоти появлення символів і біграм у російськомовних текстах, знайшла ймовірності відгадування з n-тої спроби символу, якщо відомі попередні символи тексту.