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

Криптографія

Комп’ютерний практикум No1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконав студент групи
ФБ-11 Анучін Максим

Київ - 2023

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 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. Також одержати значення H_1 та H_2 на тому ж тексті, в якому вилучено всі пробіли.

Буду наводити по 5 рядків з таблиць, ознайомитись з повним набором, можна в відповідних csv файлах

Букви:

Без пробілів: H1 = 4.467235290398882 R = 0.10655294192022358

table_letters_nospaces

Біграмма	Частота	Вірогідність
о	98914	0.10982134629241652
е	81893	0.09092342349844174
а	67731	0.07519976551076352
н	60469	0.06713697746482938
и	59977	0.06659072412985285

З пробілами H1 = 4.463449196370976 R = 0.107310160725804 79

table_letters_spaces

Біграмма	Частота	Вірогідність
	166416	0.1559520830814818
о	98914	0.09269447857130139
е	81893	0.07674372620296
а	67731	0.06347220543212098
н	60469	0.056666825977394746

Біграми без пробілів, що перетинаються:

$H_2 = 4.145794847240513$; $R = 0.5854205152759487$

table_bigrams_nospaces_cross

Біграмма	Частота	Вірогідність
ст	6662	0.014793267309144202
то	6297	0.013982768574854554
но	5561	0.012348447839410224
не	5402	0.011995381267486787
ен	5204	0.011555713460940623

Біграми без пробілів, що не перетинаються:

$H_2 = 4.1219309719166555$; $R = 0.5878069028083345$

table_bigrams_nospaces_nocross

Біграмма	Частота	Вірогідність
ст	6765	0.01514259557294523
то	6489	0.014524804534049016
но	6276	0.01404803101490085
не	5539	0.012398349871181614
ен	5061	0.011328407419759911

Біграми з пробіліми, що перетинаються:

$H_2 = 4.06477904182616$; $R = 0.593522095817384$

table_bigrams_spaces_cross

Біграмма	Частота	Вірогідність
о	10976	0.020571719882747194
е	9289	0.017409867528319852
и	8971	0.016813857422387488
с	8290	0.015537496157796487
п	8106	0.015192634964426818

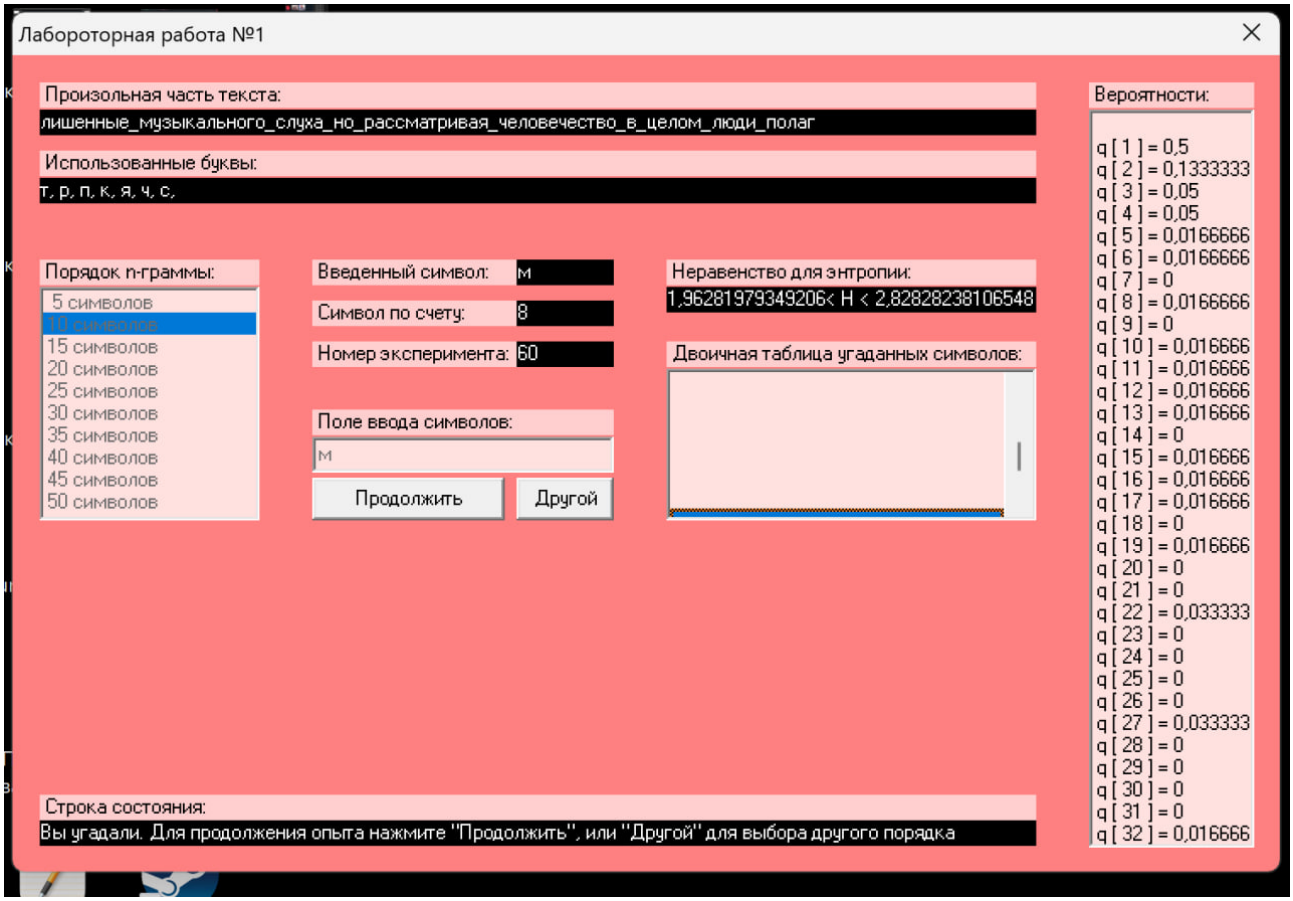
Біграми з пробілами, що не перетинаються:

$H_2 = 4.0449804653305925$; $R = 0.5955019534669408$

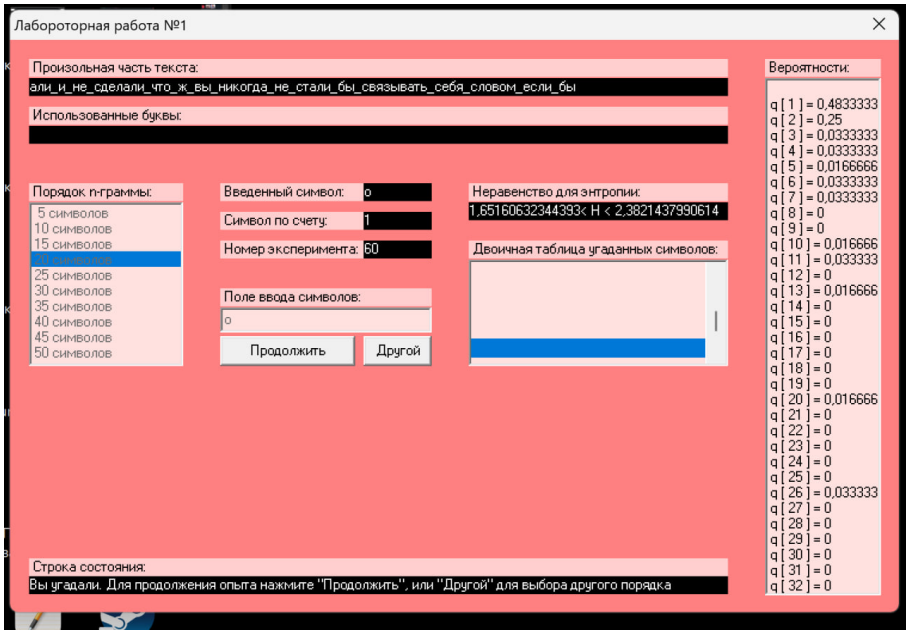
table_bigrams_spaces_nocross

Біграмма	Частота	Вірогідність
о	10662	0.020064132962988056
е	10391	0.019554155469743845
и	9502	0.017881203471610627
с	8026	0.01510361387740969
п	7855	0.014781819961008363

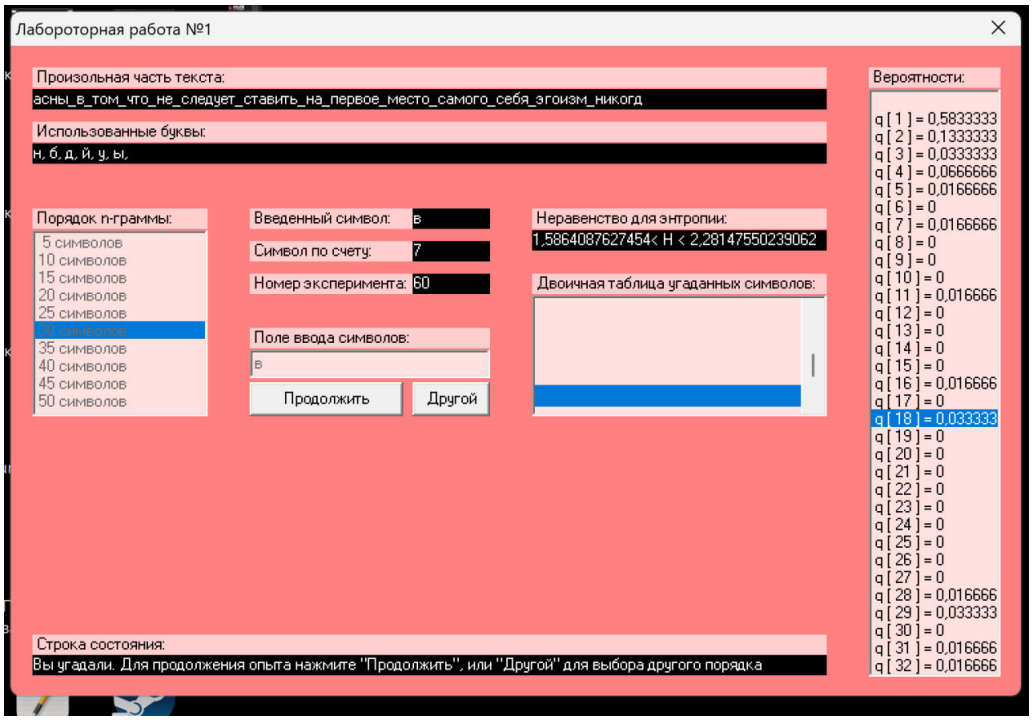
2. За допомогою програми CoolPinkProgram оцінити значення $H^{(10)}$, $H^{(20)}$, $H^{(30)}$.



1.96281979349206 H < 2.82828238106548



1.65160632344393< H < 2.3821437990614



1.5864087627454< H < 2.28147550239062

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

$R=1-H_\infty/H_0$; $H_0=\log_2 32=5$

$H^{(10)}$: 1 - 1.96281979349206/5 ≈ 0,61 < R < 1 - 2.82828238106548/5 ≈ 0,43

$H^{(20)}$: 1 - 1.65160632344393/5 ≈ 0,67 < R < 1 - 2.3821437990614/5 ≈ 0,52

$H^{(30)}$: 1 - 1.5864087627454/5 ≈ 0,68 < R < 1 - 2.28147550239062/5 ≈ 0,54

Проблеми, які виникли у ході роботи

1. Розібратись з методичкою. Виділити час на лабу). Знайти нормальний txt файл з книжкою, Фінансист Т.Д., бо багато з кривою кирилицею.

Висновки

Під час виконання лабораторної роботи ми ознайомились з поняттями ентропії на символ джерела та його надлишковості, виконали практичну перевірку за допомогою написаної власноруч програми. Поклацали CoolPinkProgram, для мануального тесту ентропії та вірогідності вгадування + можна було під час експерименту побачити, найбільш вживані біграми та в принципі найбільш вживані літери. Експериментально визначили ентропію, щоб більш наочно все перевірити.