Національний технічний університет України «Київський Політехнічний Інститут імені Ігоря Сікорського»

Комп’ютерний практикум №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконали:
студент та студентка групи ФІ-94
Маринін Іван Павло Ігорович
Немкович Ольга Михайлівна

Перевірив:
Чорний Олег Миколайович

Київ – 2022

ЗМІСТ

1. Мета 3
2. Постановка задачі 4
3. Хід роботи 5
4. Опис труднощів 6
5. Отримані частоти 7
6. Значення ентропії для відповідних частот 8
7. Результати з програми CoolPinkProgram та оцінка надлишковості 9
8. Висновки 12

1. Мета

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.

2. Постановка задачі

Оцінити значення ентропії H1 та H2 на довільному тексті російською мовою за допомогою попередньо написаної власної програми. Оцінити значення ентропії (10) H , (20) H , (30) H за допомогою програми CoolPinkProgram. Оцінити надлишковість російської мови.

3. Хід роботи

0. Уважно прочитано методичні вказівки до виконання комп’ютерного практикуму.

1. Написано програму для підрахунку частот букв і частот біграм в тексті, а також підрахунку H1 та H2 за безпосереднім означенням. Підрахувано частоти букв та біграм, а також значення H1 та H2 на обраному тексті російською мовою довжини 998 КБ, де імовірності замінені відповідними частотами. Також одержано значення H1 та H2 на тому ж тексті, в якому вилучено всі пробіли.

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

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

4. Опис труднощів

У загальному, значних труднощів при виконанні практикму не було. Виникали нюанси у написанні програмного коду, такі як робота з імпортованим текстовим файлом, проведення операцій із датафреймами та індексуваннями циків. Програмна реалізація не ідеальна, тому що потребує ручного редагування деяких невеликих ділянок коду при знаходженні частот та значень ентропії в залежності від прийняття пробілу за букву чи ні. Також отримані частоти біграм були спершу експортовані у відповідні Excel-таблиці, що потребувало подальших зусиль для внесення у звіт.

5. Отримані частоти

Таблиці частот для букв та біграм з перетинами та без для алфавіту з пробілом та без наведені в Excel-таблицях, які підписані відповідно.

6. Значення ентропії для відповідних частот

	З пробілом	Без пробілу
H_1	4.36443493250625	4.45837070945329
H_2 із перетином	3.98976701205078	4.15526780838267
H_2 без перетину	3.98817760905864	4.15556856165158

7. Результати з програми CoolPinkProgram та оцінки надлишковості

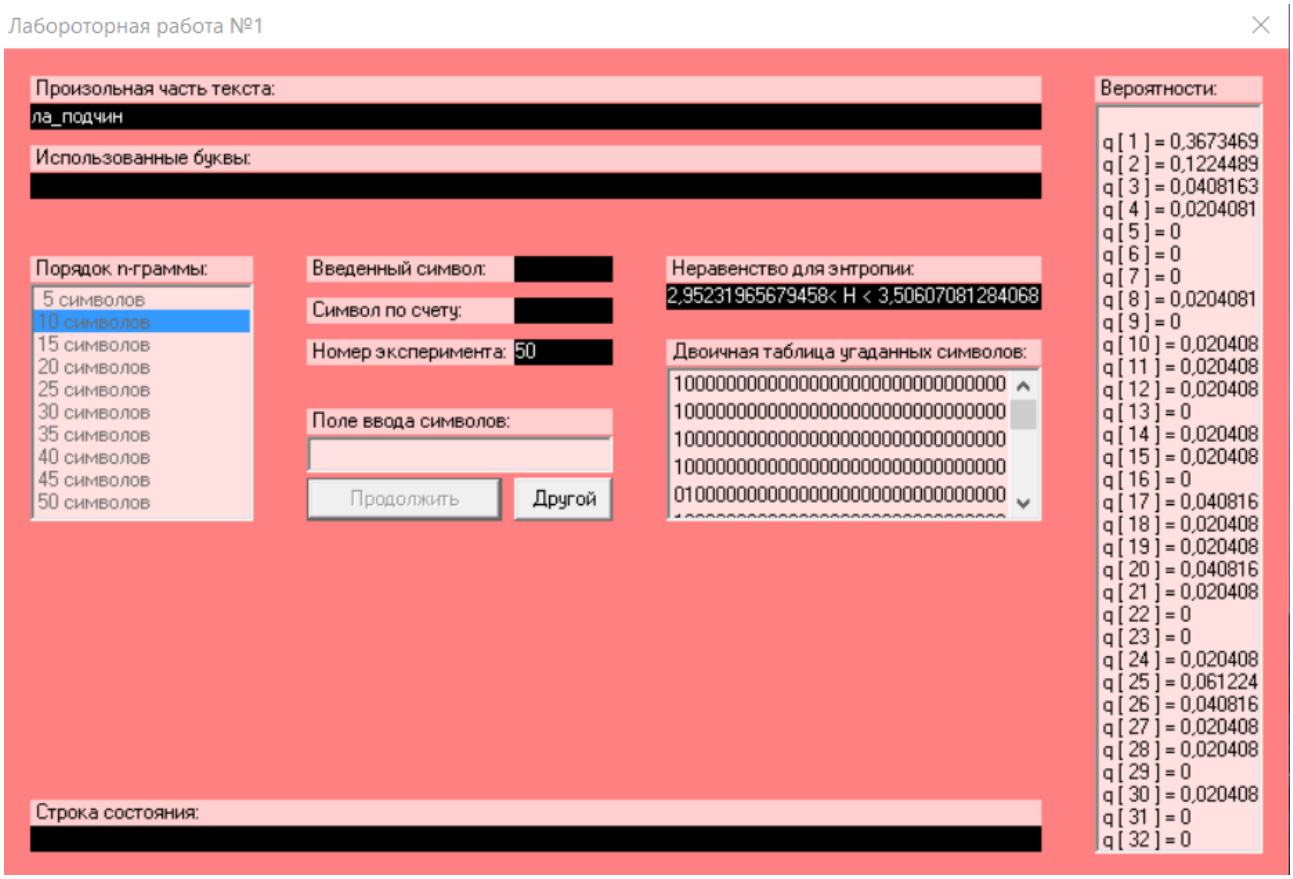

Рисунок 7.1. $H^{(10)}$

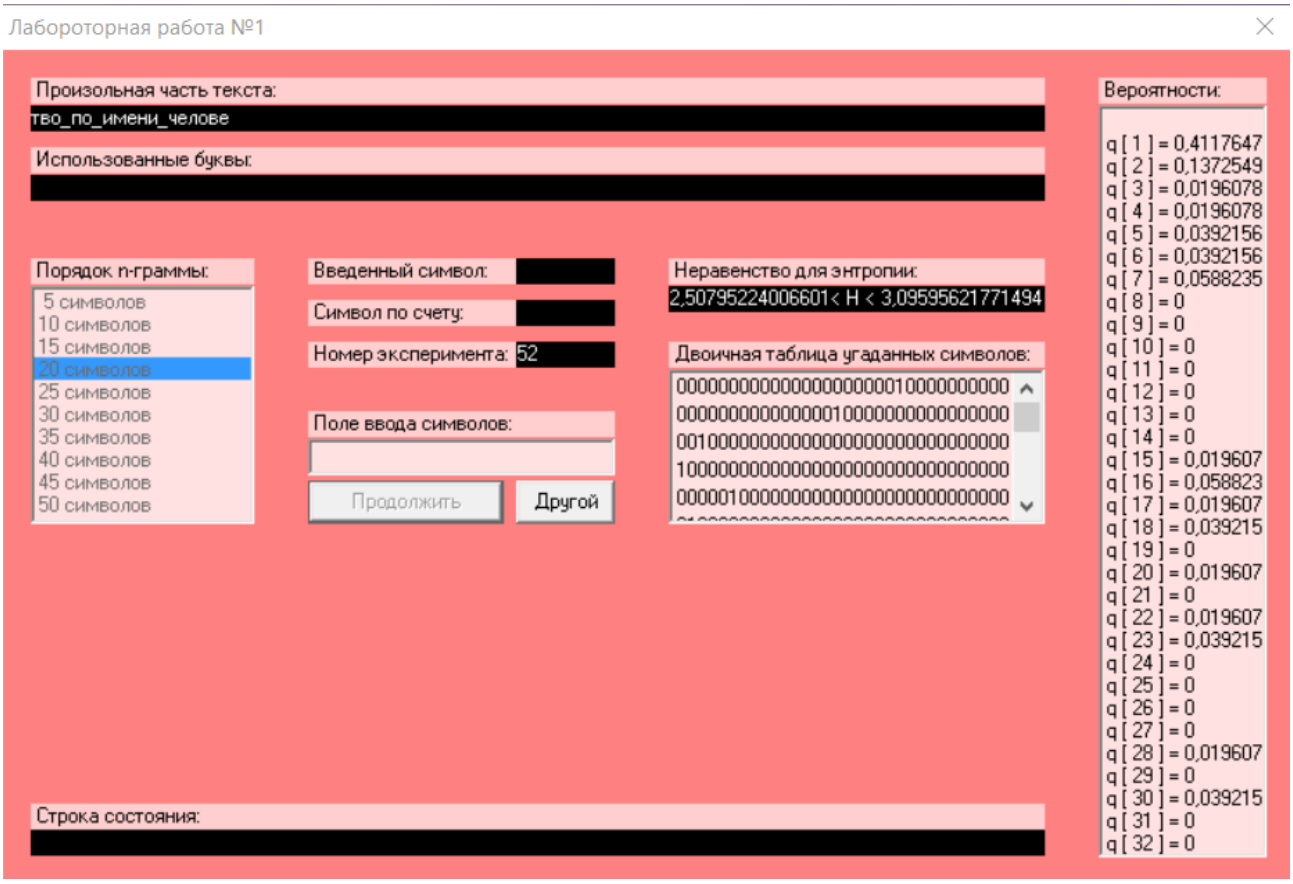

Рисунок 7.2. $H^{(20)}$

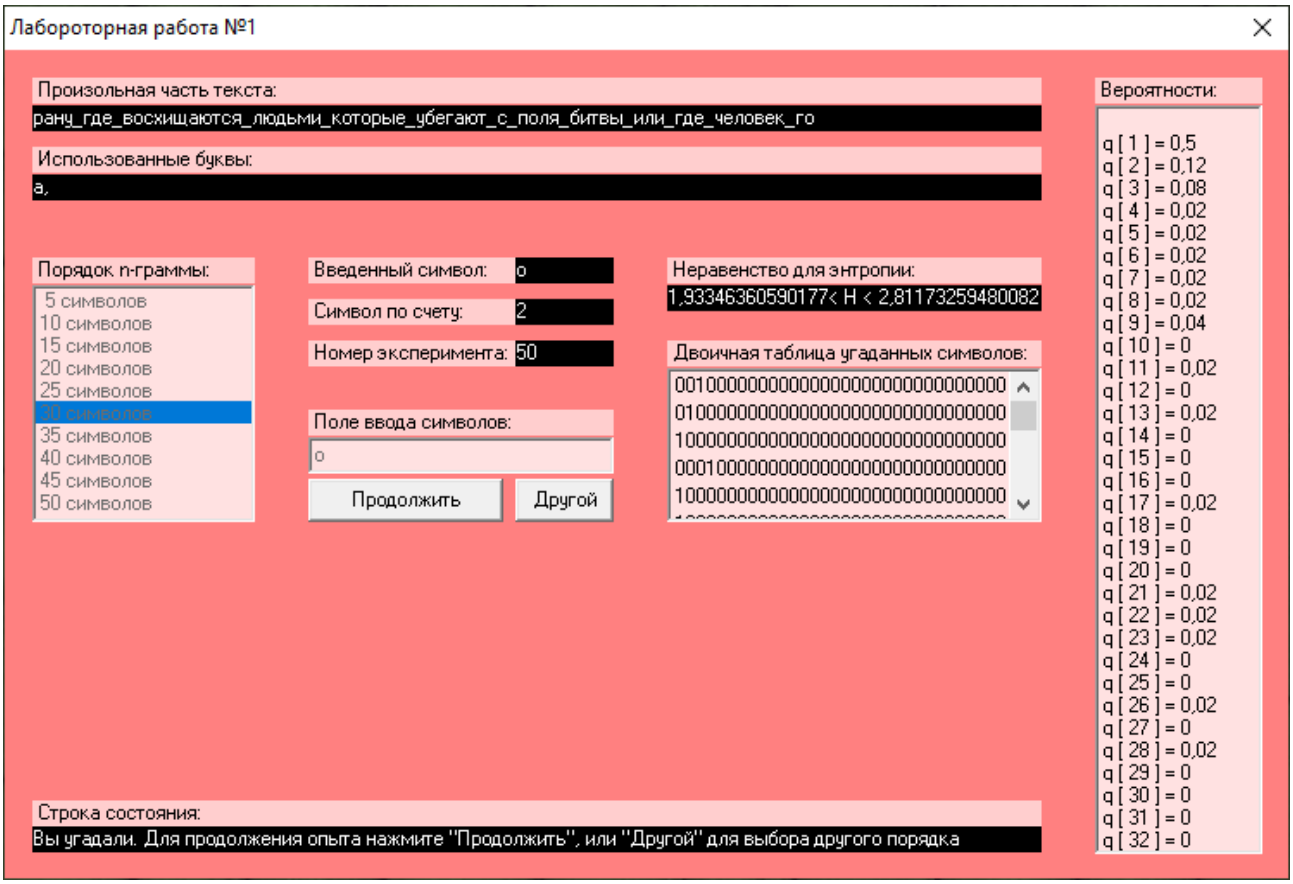

Рисунок 7.3. $H^{(30)}$

Оцінки надлишковості:

Загальна формула: $R = 1 – H_\infty/H_0$. $H_0 = \log_2 32 = 5$ для всіх моделей джерела.

- При H_1: $H_\infty = 4.36443493250625$
  $R = 1 - 4.36443493250625/5 = 1 - 0.8728869865 = 0.12711301349$

- При H_2: $H_\infty = 3.98976701205078$
  $R = 1 - 3.98976701205078/5 = 1 - 0.79795340241 = 0.20204659759$

- При $H^{(10)}$: $H_\infty = 3.22919523482$
  $R = 1 - 3.22919523482/5 = 1 - 0.64583904696 = 0.35416095303$

- При $H^{(20)}$: $H_\infty = 2.80195422889$
  $R = 1 - 2.80195422889/5 = 1 - 0.56039084577 = 0.43960915422$

- При $H^{(30)}$: $H_\infty = 2.37259810035$

  $R = 1 - 2.37259810035/5 = 1 - 0.47451962007 = 0.52548037993$

	R
H_1	0.12711301349
H_2	0.20204659759
$H^{(10)}$	0.35416095303
$H^{(20)}$	0.43960915422
$H^{(30)}$	0.52548037993

Таблиця 7.1. Значення надлишковості

8. Висновки

Під час створення даної лабораторної роботи ми обраховували частоти букв та біграм у тексті і відповідно їх ентропію. Ентропія без пробіла є більшою, ніж ентропія з пробілом (як і очікувалось, через те, що знак пробіла часто зустрічається у тексті). Також ми обрахували надлишковість джерела відкритого тексту, яка характеризує величину можливого ущільнення тексту деякою схемою кодування символів без втрати його змісту.