Національний технічний університет України «КПІ» імені Ігоря Сікорського
Фізико-технічний інститут

Комп’ютерний практикум 1
Криптографія

Виконали:
студенти ФБ-21
Князян Кирило Андрійович
Новіцький Олександр Костянтинович

Експериментальна оцінка ентропії на символ джерела відкритого тексту

Мета роботи

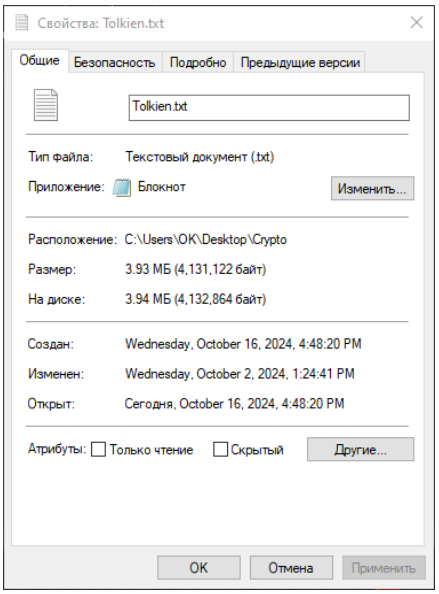
Засвоєння понять ентропії на символ джерела та його надлишковості, вивчення та порівняння різних моделей джерела відкритого тексту для наближеного визначення ентропії, набуття практичних навичок щодо оцінки ентропії на символ джерела.

Порядок виконання роботи

1. Написати програми для підрахунку частот букв і частот біграм в тексті, а також підрахунку H1 та H2 за безпосереднім означенням.
2. 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.
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:

Спочатку вибрали відповідний текст, який будемо аналізувати в нашій програмі. Ми взяли всі книги “Володар перснів”. TXT файл з серією цих книг російською мовою зайняв майже 4 мегабайти.



Після очистки файлу від усіх символів, окрім букв російського алфавіту та пробілу, заміни букв “ё” на “е” та “ъ” на “ь”, розмір файлу зменшився до 3.7 мегабайт.

Размер:	3.76 МБ (3,953,584 байт)
На диске:	3.77 МБ (3,956,736 байт)

Без пробілів:

Размер:	3.42 МБ (3,593,070 байт)
На диске:	3.42 МБ (3,596,288 байт)

Спочатку обраховуємо частоту появи букв з пробілами:

Символ	Частота	У відсотках %
	0.1664672	16.65
о	0.0925521	9.26
е	0.0670098	6.70
а	0.0658875	6.59
и	0.0580437	5.80
н	0.0562648	5.63
т	0.0455796	4.56
л	0.0446206	4.46
с	0.0440226	4.40
р	0.0401568	4.02
в	0.0344514	3.45
д	0.0285920	2.86
м	0.0267032	2.67
к	0.0259729	2.60
у	0.0244376	2.44
п	0.0227274	2.27
г	0.0175639	1.76
ь	0.0175380	1.75
ы	0.0170512	1.71
я	0.0164430	1.64
з	0.0156250	1.56
б	0.0152033	1.52
ч	0.0105580	1.06
й	0.0086660	0.87
х	0.0085082	0.85
ж	0.0076531	0.77
ш	0.0069089	0.69
ю	0.0038537	0.39
э	0.0036932	0.37
ц	0.0025908	0.26
щ	0.0023672	0.24
ф	0.0022874	0.23

Для даного випадку питома ентропія на символ H1 = 4.38

Тепер обраховуємо частоту букв без пробілів:

Символ	Частота	У відсотках %
о	0.1110360	11.10
е	0.0803925	8.04
а	0.0790461	7.90
и	0.0696357	6.96
н	0.0675016	6.75
т	0.0546825	5.47
л	0.0535319	5.35
с	0.0528144	5.28
р	0.0481766	4.82
в	0.0413318	4.13
д	0.0343021	3.43
м	0.0320361	3.20
к	0.0311600	3.12
у	0.0293181	2.93
п	0.0272664	2.73
г	0.0210717	2.11
ь	0.0210405	2.10
ы	0.0204566	2.05
я	0.0197269	1.97
з	0.0187455	1.87
б	0.0182396	1.82
ч	0.0126666	1.27
й	0.0103967	1.04
х	0.0102074	1.02
ж	0.0091816	0.92
ш	0.0082887	0.83
ю	0.0046233	0.46
э	0.0044308	0.44
ц	0.0031082	0.31
щ	0.0028399	0.28
ф	0.0027442	0.27

Питома ентропія на символ для даного випадку H1 = 4.47

Тепер обрахуємо частоту появи найчастіших біграм з кроком 1 з пробілами. Для зручності, виведемо лише 10 біграм, що зустрічаються найчастіше:

Біграма	Частота	У відсотках %
'и '	0.0214599	2.15
'о '	0.0208191	2.08
'е '	0.0173672	1.74
' н'	0.0170373	1.70
' с'	0.0161526	1.62
' п'	0.0159897	1.60
'а '	0.0158134	1.58
' в'	0.0150427	1.50
' о'	0.0116776	1.17
' и'	0.0115945	1.16

Питома ентропія на символ H2 = 3.98

Частота появи 10 найчастіших біграм з кроком 2 з пробілами:

Біграма	Частота	У відсотках %
'и '	0.0214334	2.14
'о '	0.0207536	2.08
'е '	0.0174220	1.74
' н'	0.0170885	1.71
' с'	0.0161176	1.61
' п'	0.0159731	1.60
'а '	0.0157968	1.58
' в'	0.0151173	1.51
' о'	0.0116802	1.17
' и'	0.0116242	1.16

Питома ентропія на символ H2 = 3.98

Частота появи 10 найчастіших біграм з кроком 1 без пробілів:

Біграма	Частота	У відсотках %
то	0.0125464	1.25
на	0.0120510	1.21
не	0.0118194	1.18
ал	0.0116753	1.17
ли	0.0114487	1.14
ст	0.0113274	1.13
но	0.0111971	1.12
по	0.0111431	1.11
он	0.0106238	1.06
ро	0.0100839	1.01

Питома ентропія на символ H2 = 4.16

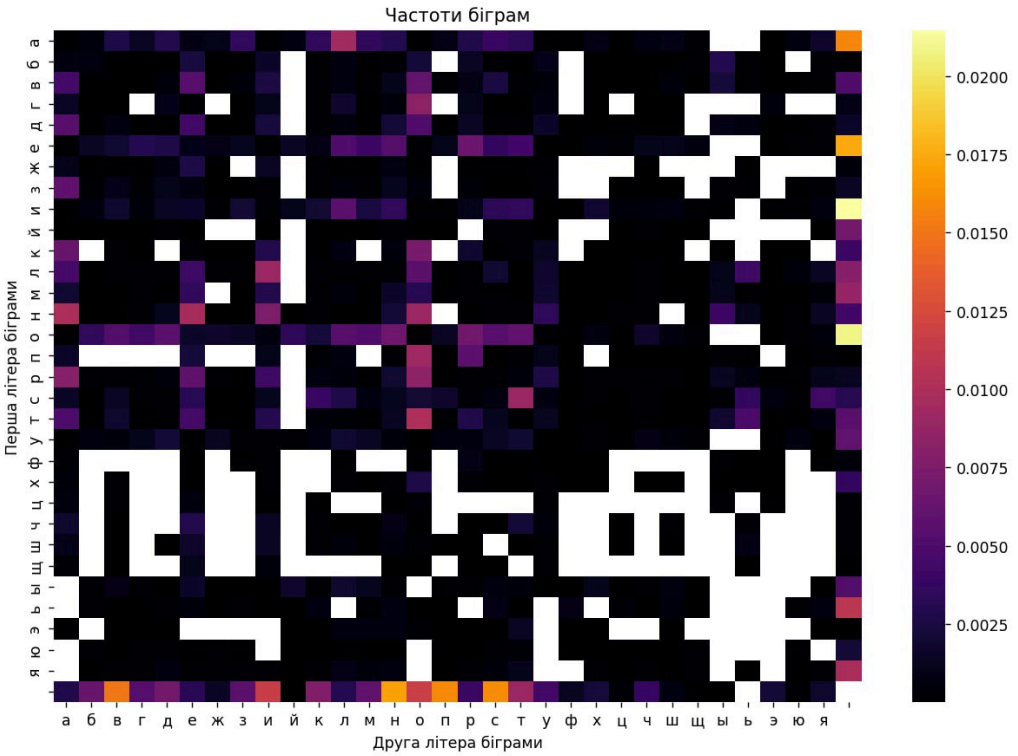
Частота появи 10 найчастіших біграм з кроком 2 без пробілів:

Біграма	Частота	У відсотках %
то	0.0125468	1.25
на	0.0120992	1.21
не	0.0118280	1.18
ал	0.0116539	1.17
ли	0.0114354	1.14
ст	0.0113574	1.14
но	0.0112146	1.12
по	0.0111340	1.11
он	0.0105993	1.06
ро	0.0100783	1.01

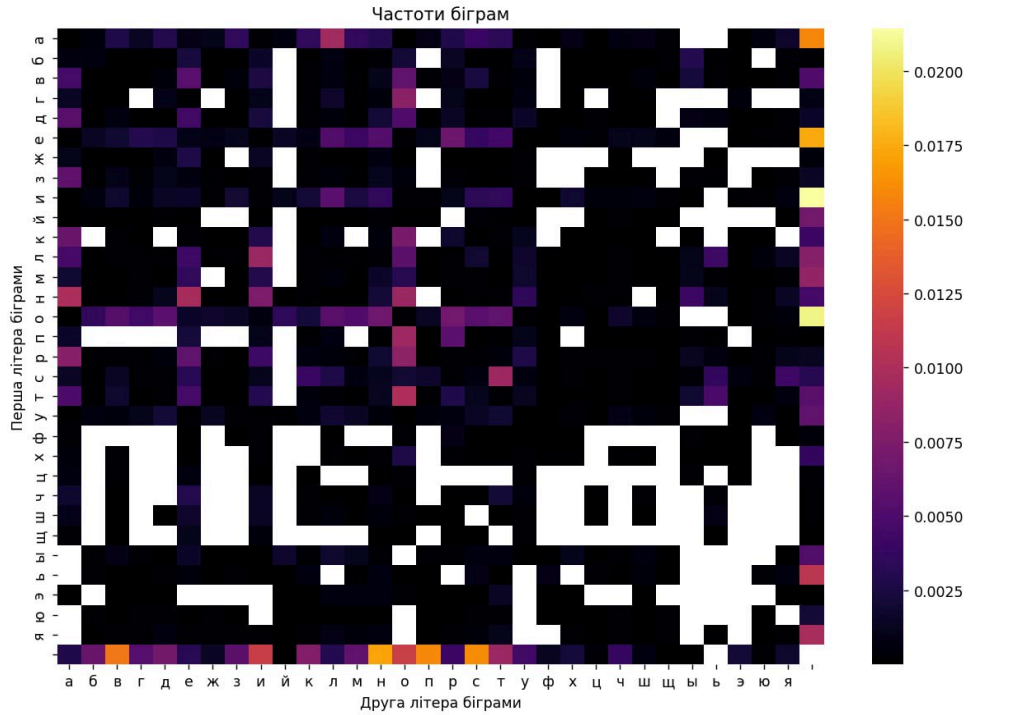
Питома ентропія на символ H2 = 4.16

Тепер, побудуємо heatmaps для біграм, задля зручної візуалізації частоти їх появи в тексті.

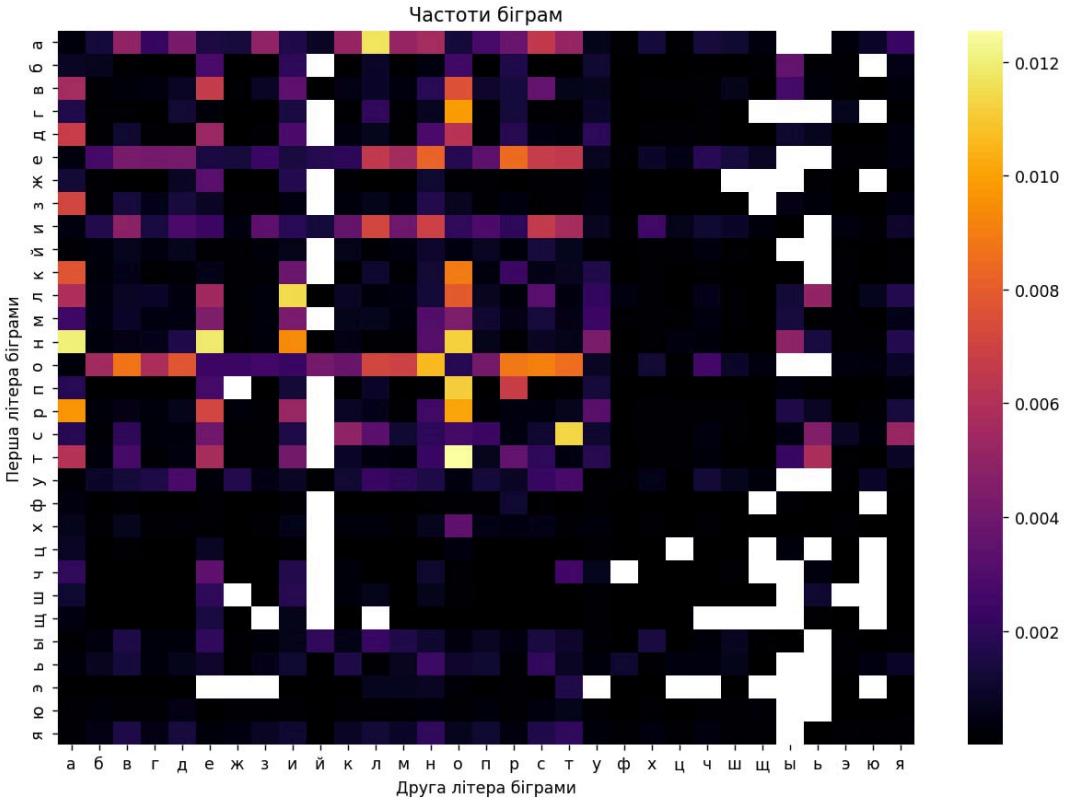
Для біграм в тексті з пробілами з кроком 1:



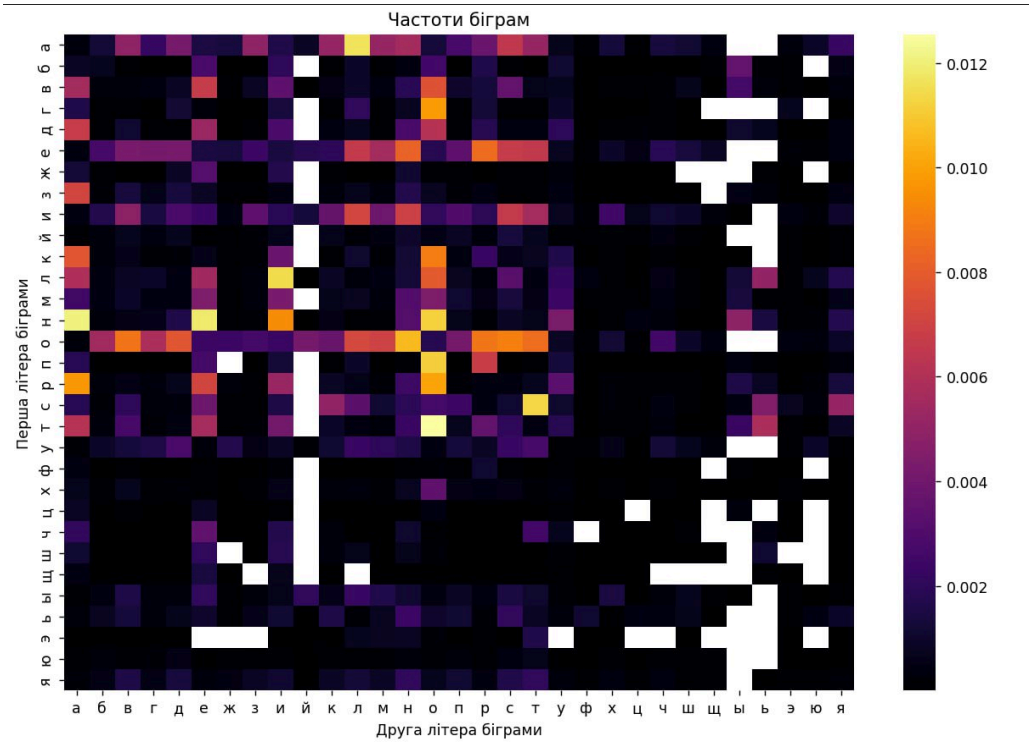
Для біграм в тексті з пробілами з кроком 2:



Для біграм в тексті без пробілів з кроком 1:

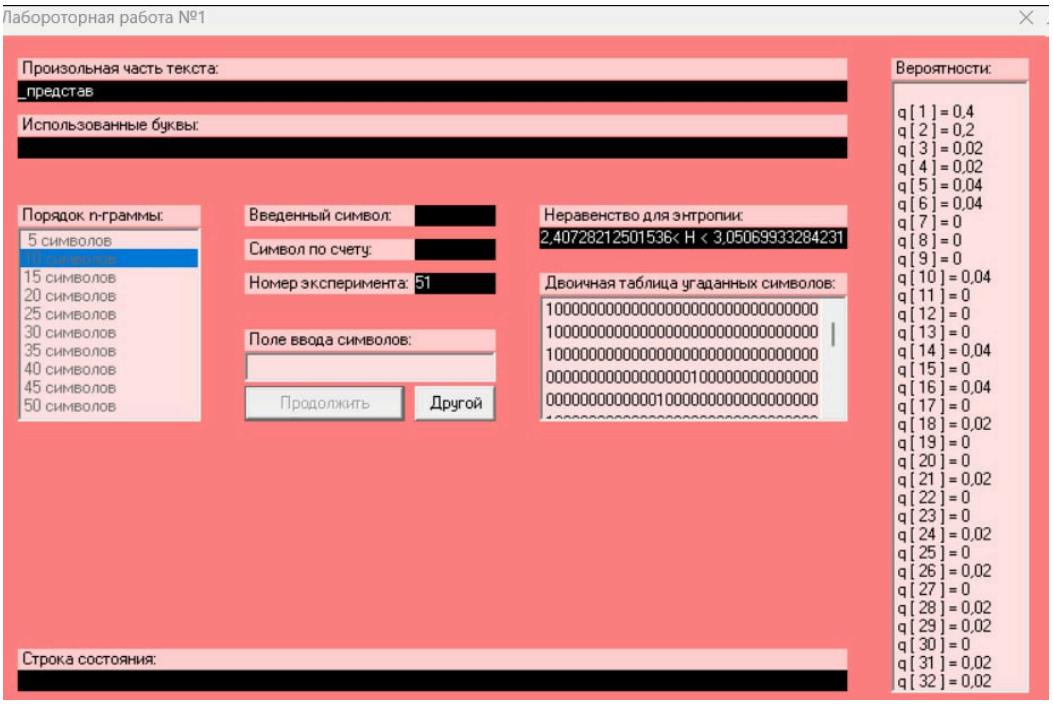


Для біграм в тексті без пробілів з кроком 2:

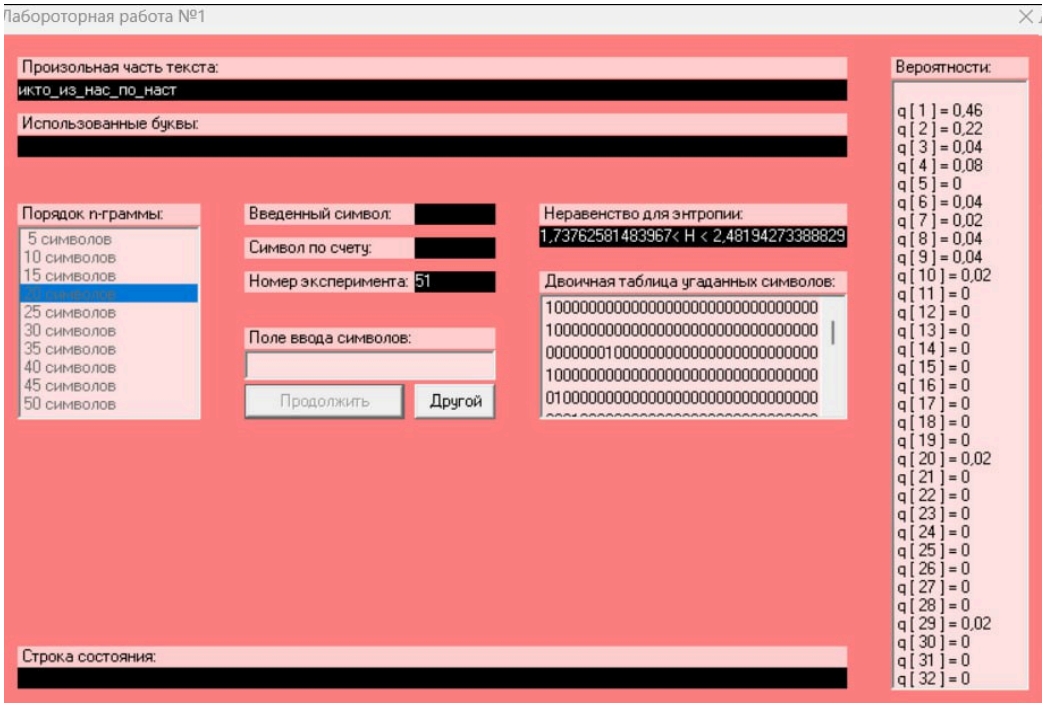


Тепер перейдемо до програми CoolPinkProgram. Спочатку вираховуємо

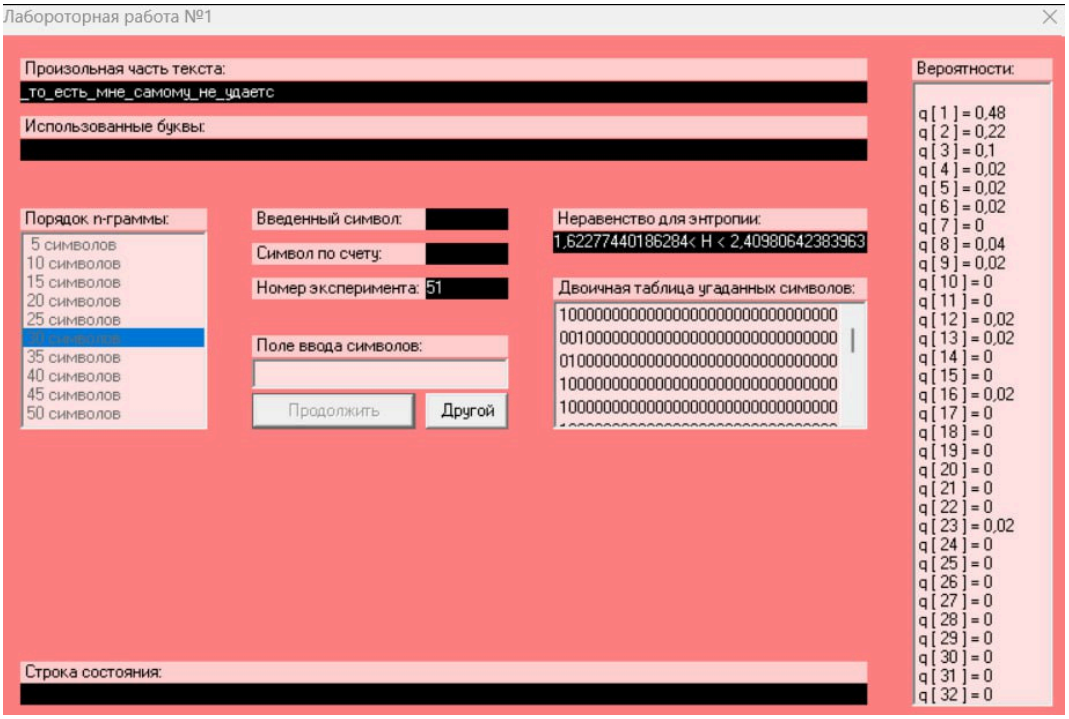
H(10):


Лабороторная работа №1
Произольная часть текста:
_представ
Использованные буквы:
Порядок n-граммы:
5 символов
10 символов
15 символов
20 символов
25 символов
30 символов
35 символов
40 символов
45 символов
50 символов
Введенный символ:
Символ по счету:
Номер эксперимента: 51
Поле ввода символов:
Продолжить
Другой
Неравенство для энтропии:
2,40728212501536< H < 3,05069933284231
Двоичная таблица угаданных символов:
Строка состояния:
Вероятности:
q [1] = 0,4
q [2] = 0,2
q [3] = 0,02
q [4] = 0,02
q [5] = 0,04
q [6] = 0,04
q [7] = 0
q [8] = 0
q [9] = 0
q [10] = 0,04
q [11] = 0
q [12] = 0
q [13] = 0
q [14] = 0,04
q [15] = 0
q [16] = 0,04
q [17] = 0
q [18] = 0,02
q [19] = 0
q [20] = 0
q [21] = 0,02
q [22] = 0
q [23] = 0
q [24] = 0,02
q [25] = 0
q [26] = 0,02
q [27] = 0
q [28] = 0,02
q [29] = 0,02
q [30] = 0
q [31] = 0,02
q [32] = 0,02

H(20):


Лабороторная работа №1
Произольная часть текста:
икто_из_нас_по_наст
Использованные буквы:
Порядок n-граммы:
5 символов
10 символов
15 символов
20 символов
25 символов
30 символов
35 символов
40 символов
45 символов
50 символов
Введенный символ:
Символ по счету:
Номер эксперимента: 51
Поле ввода символов:
Продолжить
Другой
Неравенство для энтропии:
1,73762581483967< H < 2,48194273388829
Двоичная таблица угаданных символов:
Строка состояния:
Вероятности:
q [1] = 0,46
q [2] = 0,22
q [3] = 0,04
q [4] = 0,08
q [5] = 0
q [6] = 0,04
q [7] = 0,02
q [8] = 0,04
q [9] = 0,04
q [10] = 0,02
q [11] = 0
q [12] = 0
q [13] = 0
q [14] = 0
q [15] = 0
q [16] = 0
q [17] = 0
q [18] = 0
q [19] = 0
q [20] = 0,02
q [21] = 0
q [22] = 0
q [23] = 0
q [24] = 0
q [25] = 0
q [26] = 0
q [27] = 0
q [28] = 0
q [29] = 0,02
q [30] = 0
q [31] = 0
q [32] = 0

H(30):



За допомогою формули обрахуємо надлишковість для всіх отриманих ентропій:

Питома ентропія на символ H2 з перетинами біграм і без вийшло однаково з такою точністю, тому обрахуємо по 1 разу:

Текст з пробілами:

$$H0 = \ln 32 = 5$$

H1 = 4.38 R = 1 - 4.38 / 5 = 0.124

H2 = 3.98 R = 1 - 3.98 / 5 = 0.204

Текст без пробілів:

$$H0 = \ln 31 = 4.95$$

H1 = 4.47 R = 1 - 4.47 / 4.95 = 0.096

H2 = 4.16 R = 1 - 4.16 / 4.95 = 0.159

CoolPinkProgram:

$H0 = \ln 32 = 5$

$2.41 < H(10) < 3.05$	$0,518 > R > 0,39$
$1.74 < H(20) < 2.48$	$0,652 > R > 0,504$
$1.62 < H(30) < 2.41$	$0,676 > R > 0,518$

Висновки:

У ході виконання лабораторної роботи, ми навчились писати python код для обрахунку частоти появи конкретних символів та біграм у тексті. Ми виконували обрахунки на тексті з пробілами та без пробілів. Також ми вирахували питому ентропію для кожного з випадків, використовуючи формули, які ми дізнались з теорії. В кінці, ми провели експеримент з програмою CoolPinkProgram, що допомогло нам експериментальним шляхом визначити залежність ентропії та надлишковості від кількості символів.