КРИПТОГРАФІЯ

КОМП’ЮТЕРНИЙ ПРАКТИКУМ №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконали: ФБ-11 Мельниченко Богдан, Захаренко Нікіта

Варіант: 8

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:

0. Уважно прочитати методичні вказівки до виконання комп’ютерного практикуму.
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:

1.

Для роботи над практикумом було об’єднано дві книги “Первая експедиция” Антон Первушин та “Коготь химеры” Сергей Павловский, обраний текст знаходиться у файлі **text.txt(1,59 МБ).**

Для фільтрування тексту від зайвих символів була написана програма **text_redacher_spaces.py**, вивід знаходиться у файлі **spaces_text.txt**

text_redacher_spaces.py > ...

```python
alphabet = 'абвгдежзийклмнопрстуфхцчшщыьэюя '

with open( 'text.txt' , 'r', encoding='utf-8') as f:
    text = f.read().lower()

result = ''
for i in text:
    if i in alphabet:
        result += i
    if i == '\n':
        result += ' '
    if i == 'ъ':
        result += 'ь'

with open('spaces_text.txt', 'w', encoding='utf-8') as nf:
    nf.write(" ".join(result.split()))
```

Для фільтрування тексту від зайвих символів та пробілів була написана програма **text_redacher_nospaces.py**, вивід знаходиться у файлі **NOspaces_text.txt**

text_redacher_nospaces.py > ...

```python
alphabet = 'абвгдежзийклмнопрстуфхцчшщыьэюя'

with open( 'text.txt' , 'r', encoding='utf-8') as f:
    text = f.read().lower()

result = ''
for i in text:
    if i in alphabet:
        result += i
    if i == 'ь':
        result += 'ь'


with open('NOspaces_text.txt', 'w', encoding='utf-8') as nf:
    nf.write(" ".join(result.split()))
```

Частота букв у тексті з пробілами та без них**(H1_results.py)**

[(' ', 132030),
('о', 82127),
('е', 59219),
('а', 57937),
('н', 49497),
('и', 48651),
('т', 43940),
('с', 37807),
('л', 36385),
('р', 35143),
('в', 31538),
('к', 27547),
('м', 23373),
('д', 22659),
('п', 22072),
('у', 21818),
('ы', 14331),
('я', 14329),
('з', 13444),
('ь', 12936),
('б', 12167),
('г', 11551),
('ч', 10100),
('й', 8714),
('х', 6933),
('ж', 6915),
('ш', 5778),
('ю', 4813),
('ц', 3023),
('щ', 2328),
('э', 1964),
('ф', 1518)]
H1 (spaces): 4.3966511464179385

[('о', 82127),
('е', 59219),
('а', 57937),
('н', 49497),
('и', 48651),
('т', 43940),
('с', 37807),
('л', 36385),
('р', 35143),
('в', 31538),
('к', 27547),
('м', 23373),
('д', 22659),
('п', 22072),
('у', 21818),
('ы', 14331),
('я', 14329),
('з', 13444),
('ь', 12936),
('б', 12167),
('г', 11551),
('ч', 10100),
('й', 8714),
('х', 6933),
('ж', 6915),
('ш', 5778),
('ю', 4813),
('ц', 3023),
('щ', 2328),
('э', 1964),
('ф', 1518),
(' ', 0)]
H1 (no spaces): 4.462195077627556

Частота біграм(**H2_results.py**)
У тексті:

Без перетину, без пробілів
('но', 5135),
('то', 4965),
('ст', 4687),
('по', 4234),
('на', 4170),
('ов', 4013),
('ро', 3741),
('ко', 3720),
('ал', 3717),
('не', 3635),
('ен', 3613),
('ра', 3540),
('он', 3373),
('ер', 3363),
('ос', 3281),

…
('юы', 0),
('юь', 0),
('яы', 0),
('яь', 0)
H2: 4.164582058626493

З перетином, без пробілів
('но', 10244),
('то', 10001),
('ст', 9301),
('по', 8421),
('на', 8399),
('ов', 8073),
('ал', 7488),
('ро', 7378),
('ко', 7348),
('не', 7269),
('ен', 7170),
('ра', 7130),
('он', 6808),
('ер', 6666),
('ос', 6590),
…
('юы', 0),
('юь', 0),
('яы', 0),
('яь', 0)
H2: 4.165162619108788

Без перетину, з пробілами
('но', 5034),
('то', 4922),
('ст', 4516),
('на', 4207),
('по', 4207),
('не', 3632),
('ал', 3630),
('ро', 3599),
('ра', 3490),
('ко', 3482),
('ов', 3208),
('ер', 3171),
('ен', 2995),
('ни', 2943),
('пр', 2924),
…
('юя', 0),
('яф', 0),
('яы', 0),
('яь', 0)
H2: 2.9610851668557623

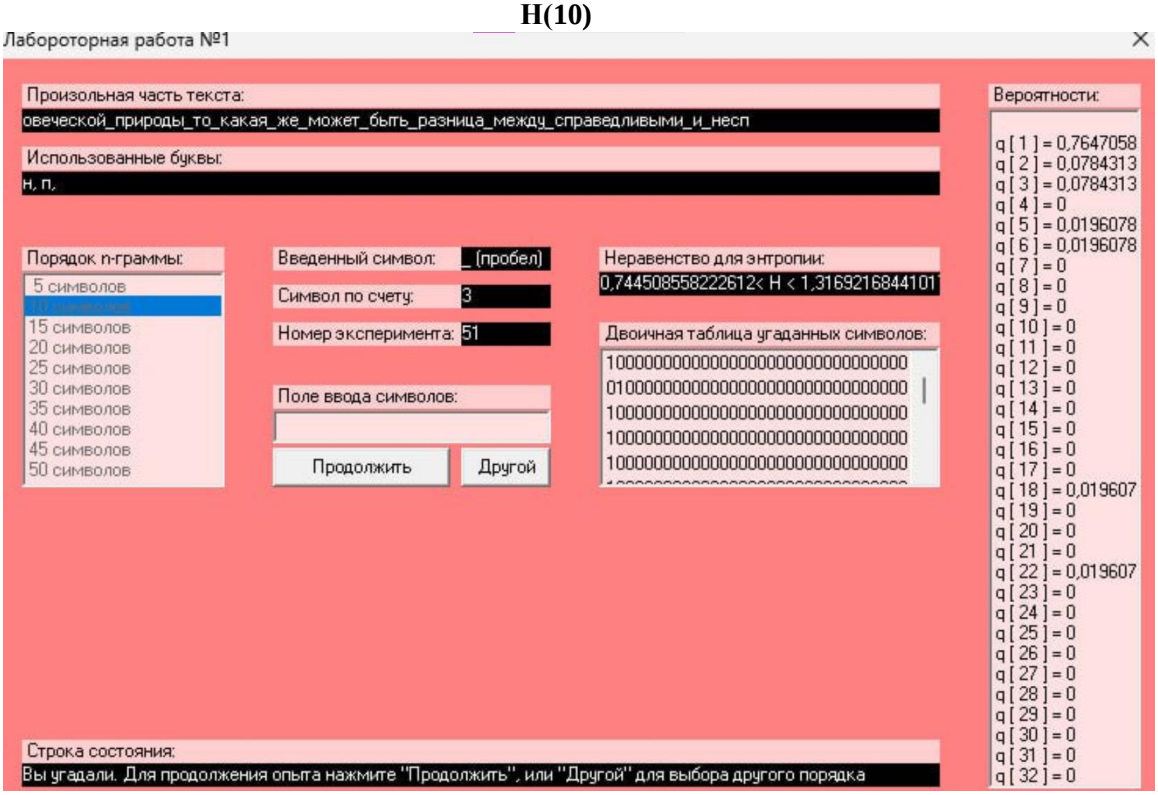
З перетином, з пробілами

('но', 10143),
('то', 9744),
('ст', 9131),
('по', 8415),
('на', 8357),
('ро', 7313),
('ал', 7269),
('не', 7228),
('ра', 7102),
('ко', 7046),
('ов', 6478),
('ер', 6281),
('ен', 5938),
('пр', 5854),
('ни', 5840),
…
('юя', 0),
('яф', 0),
('яы', 0),
('яь', 0)]

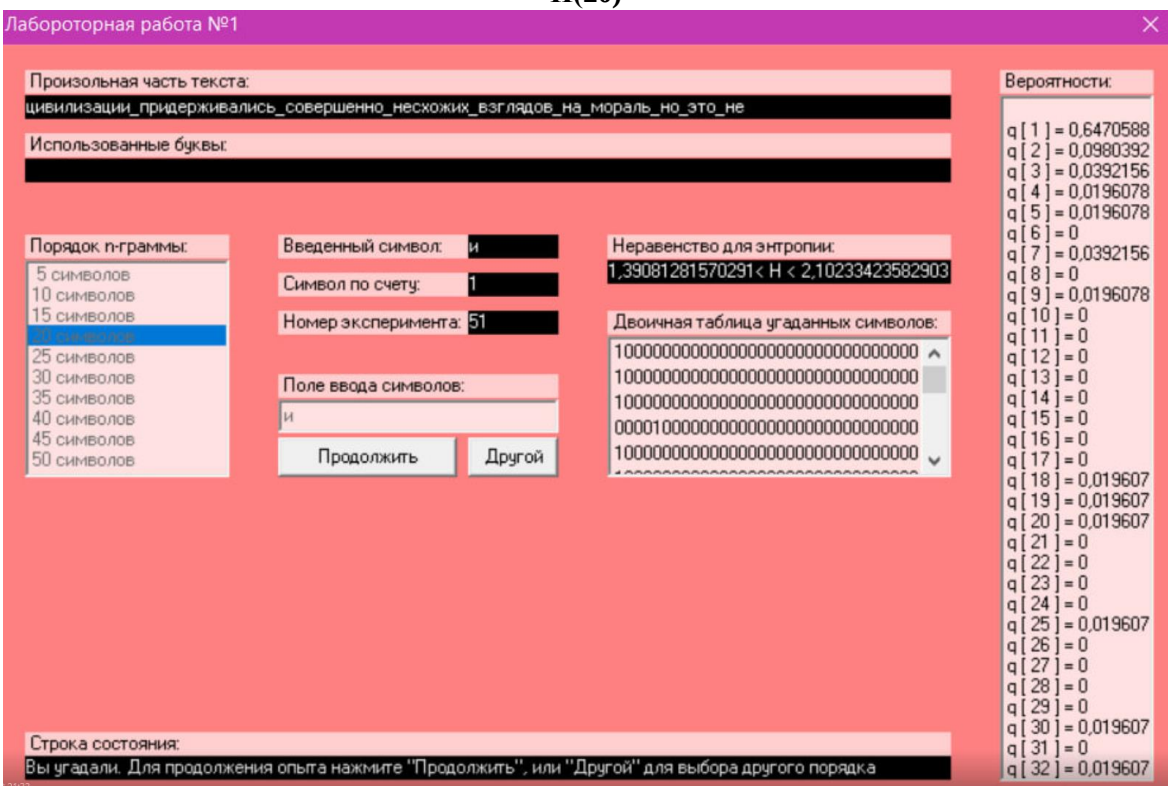
H2: 2.9617338550695043

2.

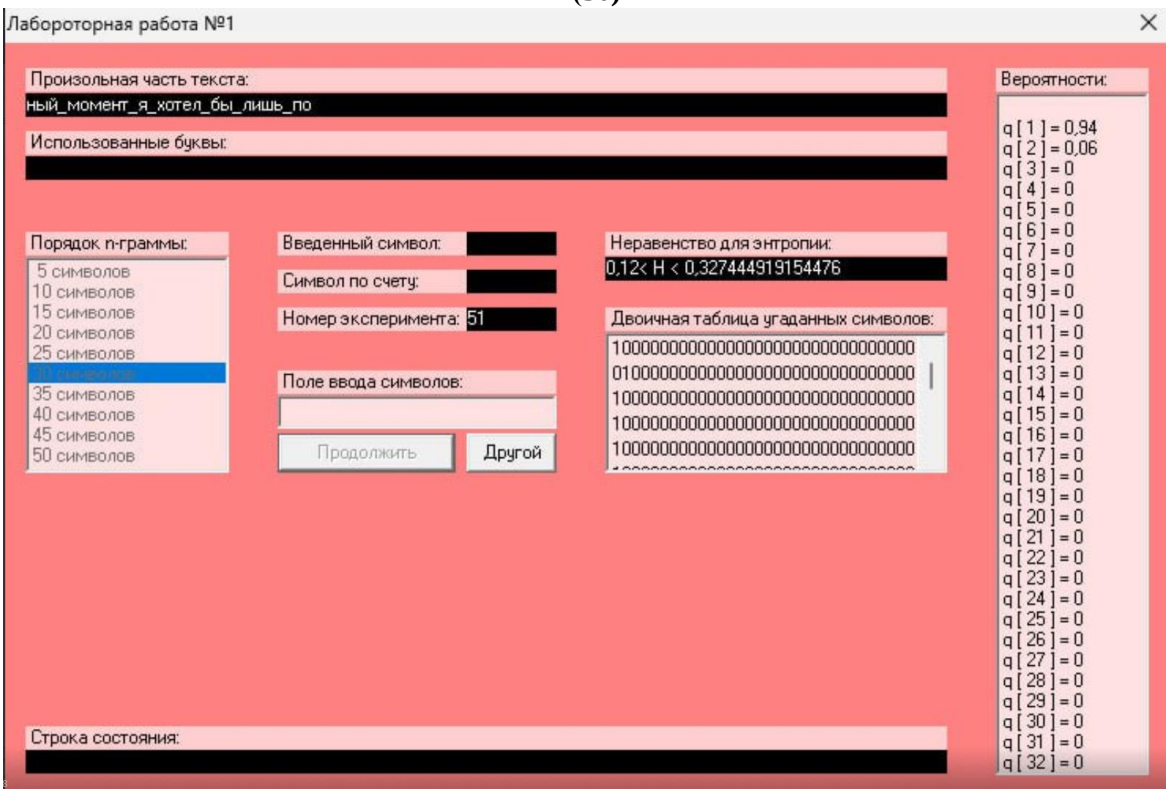
H(10)



H(20)


Лабороторная работа №1
Произольная часть текста:
цивилизации_придерживались_совершенно_несхожих_взглядов_на_мораль_но_это_не
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: и
Символ по счету: 1
Номер эксперимента: 51
Поле ввода символов:
и
Продолжить
Другой
Неравенство для энтропии:
1,39081281570291< H < 2,10233423582903
Двоичная таблица угаданных символов:
Строка состояния:
Вы угадали. Для продолжения опыта нажмите "Продолжить", или "Другой" для выбора другого порядка
Вероятности:
q[1] = 0,6470588
q[2] = 0,0980392
q[3] = 0,0392156
q[4] = 0,0196078
q[5] = 0,0196078
q[6] = 0
q[7] = 0,0392156
q[8] = 0
q[9] = 0,0196078
q[10] = 0
q[11] = 0
q[12] = 0
q[13] = 0
q[14] = 0
q[15] = 0
q[16] = 0
q[17] = 0
q[18] = 0,019607
q[19] = 0,019607
q[20] = 0,019607
q[21] = 0
q[22] = 0
q[23] = 0
q[24] = 0
q[25] = 0,019607
q[26] = 0
q[27] = 0
q[28] = 0
q[29] = 0
q[30] = 0,019607
q[31] = 0
q[32] = 0,019607

H(30)


Лабороторная работа №1
Произольная часть текста:
ный_момент_я_хотел_бы_лишь_по
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:
0,12< H < 0,327444919154476
Двоичная таблица угаданных символов:
Строка состояния:
Вероятности:
q[1] = 0,94
q[2] = 0,06
q[3] = 0
q[4] = 0
q[5] = 0
q[6] = 0
q[7] = 0
q[8] = 0
q[9] = 0
q[10] = 0
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
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
q[30] = 0
q[31] = 0
q[32] = 0

3.

З методички:

Надлишковість джерела відкритого тексту (мови) дорівнює $R = 1 - \frac{H_\infty}{H_0}$

де $H_0 = \log_2 m$ (m – кількість символів алфавіту). R називають також надлишковістю джерела.

$H_0 = \log_2 32 = 5$ (для тексту з пробілами)
$H_0 = \log_2 31 = 4{,}95$(для тексту без пробілів)

Надлишковість для TEXT:

H(10): 73.8% < R < 85.2%
H(20): 58% < R < 72.2%
H(30): 93.4% < R < 97.6%

Надлишковість для тексту з книг:

H1 (spaces): 4.3966511464179385
R = 12,2%
H1 (no spaces): 4.462195077627556
R = 9,89%
Без перетину, без пробілів
H2: 4.164582058626493
R = 15,95%
З перетином, без пробілів
H2: 4.165162619108788
R = 15,95%
Без перетину, з пробілами
H2: 2.9610851668557623
R = 40,8%
З перетином, з пробілами
H2: 2.9617338550695043
R = 40,8%

Висновок: під час виконання цієї лабораторної роботи, ми освоїли концепції ентропії та надлишковості. Ми створили кілька програм на за допомогою Python, що обчислюють частоту кожної букви в тексті, а також частоту біграм і H1 та H2 за означенням. На основі отриманих даних ми розрахували значення ентропії. Після застосування програми CoolPinkProgram.exe, ми визначили межі умовної ентропії джерела. Зрештою, ми оцінили надлишковість російської мови в різних моделях джерела.