Міністерство освіти і науки України

Національний технічний університет України  
«Київський політехнічний інститут

імені ІГОРЯ СІКОРСЬКОГО»

Компьютерний практикум №1

З дисципліни: «Криптографія»

Виконав:

Студент гр. ФБ-03

Гузенков А.М.

Перевірив:

Чорний О.М.

Київ – 2022

**Тема**

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. 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 H(10), H(20), H(30).

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

**Хід роботи**

Завдання 1

Для виконання першого завдання був написаний код (див. файл main.cpp), який підраховує частоту літер та біграм у тексті та рахує ентропію.

Таблиця з частотою літер:

|  |  |  |
| --- | --- | --- |
| Літера | Частота | Ймовірність |
| а | 13 536 | 0,0782071 |
| б | 3 029 | 0,0175007 |
| в | 8 379 | 0,0484114 |
| г | 3 362 | 0,0194247 |
| д | 4 858 | 0,0280681 |
| е | 14 979 | 0,0865443 |
| ж | 1 841 | 0,0106368 |
| з | 3 015 | 0,0174198 |
| и | 12 162 | 0,0702685 |
| й | 2 419 | 0,0139763 |
| к | 5 533 | 0,0319681 |
| л | 8 060 | 0,0465683 |
| м | 5 229 | 0,0302116 |
| н | 12 496 | 0,0721982 |
| о | 19 720 | 0,113936 |
| п | 4 362 | 0,0252024 |
| р | 7 552 | 0,0436333 |
| с | 9 552 | 0,0551887 |
| т | 9 920 | 0,0573149 |
| у | 4 448 | 0,0256992 |
| ф | 284 | 0,00164087 |
| х | 1 644 | 0,00949855 |
| ц | 661 | 0,00381907 |
| ч | 2 349 | 0,0135718 |
| ш | 1 213 | 0,00700836 |
| щ | 576 | 0,00332796 |
| ъ | 46 | 0,000265775 |
| ы | 3 270 | 0,0188931 |
| ь | 3 109 | 0,0179629 |
| э | 522 | 0,00301596 |
| ю | 1 190 | 0,00687547 |
| я | 3 760 | 0,0217242 |
| ё | 3 | 0,0000173331 |

Таблиця з частотою літер з пробілами:

|  |  |  |
| --- | --- | --- |
| Биграмма | Частота | Вероятность |
|  | 31 299 | 0,153143 |
| а | 13 536 | 0,0662302 |
| б | 3 029 | 0,0148206 |
| в | 8 379 | 0,0409976 |
| г | 3 362 | 0,0164499 |
| д | 4 858 | 0,0237697 |
| е | 14 979 | 0,0732907 |
| ж | 1 841 | 0,00900782 |
| з | 3 015 | 0,0147521 |
| и | 12 162 | 0,0595074 |
| й | 2 419 | 0,0118359 |
| к | 5 533 | 0,0270724 |
| л | 8 060 | 0,0394367 |
| м | 5 229 | 0,0255849 |
| н | 12 496 | 0,0611416 |
| о | 19 720 | 0,0964879 |
| п | 4 362 | 0,0213428 |
| р | 7 552 | 0,0369511 |
| с | 9 552 | 0,0467369 |
| т | 9 920 | 0,0485375 |
| у | 4 448 | 0,0217636 |
| ф | 284 | 0,00138958 |
| х | 1 644 | 0,00804392 |
| ц | 661 | 0,0032342 |
| ч | 2 349 | 0,0114934 |
| ш | 1 213 | 0,00593508 |
| щ | 576 | 0,00281831 |
| ъ | 46 | 0,000225073 |
| ы | 3 270 | 0,0159998 |
| ь | 3 109 | 0,015212 |
| э | 522 | 0,00255409 |
| ю | 1 190 | 0,00582254 |
| я | 3 760 | 0,0183973 |
| ё | 3 | 0,0000146787 |

Таблиця з частотою біграм без пробілів та без перетинів (перші 10):

|  |  |  |
| --- | --- | --- |
| Біграма | Частота | Ймовірність |
| аа | 38 | 0,000439103 |
| аб | 132 | 0,00152531 |
| ав | 511 | 0,00590478 |
| аг | 102 | 0,00117865 |
| ад | 165 | 0,00190663 |
| ае | 141 | 0,0016293 |
| аж | 185 | 0,00213774 |
| аз | 430 | 0,0049688 |
| аи | 144 | 0,00166397 |
| ай | 46 | 0,000531546 |

Переглянути повні таблиці частот біграм та монограм можна у файлах

table\_bigrams\_nospaces\_cross.csv — біграми без пробілів та з перетинами

table\_bigrams\_nospaces\_nocross.csv — біграми без пробілів та без перетинів

table\_bigrams\_spaces\_cross.csv — біграми з пробілами та з перетинами

table\_bigrams\_spaces\_nocross.csv — біграми з пробілами та без перетинів

table\_letters\_nospaces.csv — монограми без пробілів

table\_letters\_spaces.csv — монограми без пробілів

Результат роботи програми можна побачити нижче

Ентропія монограм без пробілів: 4,45249

Надлишковість для тексту без пробілів: 0,109503

Ентропія монограм з пробілами: 4,38827

Надлишковість для тексту з пробілами: 0,122346

Ентропія біграм без пробілів та з перетинами: 4,12868

Надлишковість для джерела біграм без пробілів та з перетинами: 0,174265

Ентропія біграм без пробілів та без перетинів: 4,10195

Надлишковість для джерела біграм без пробілів та без перетинів: 0,179611

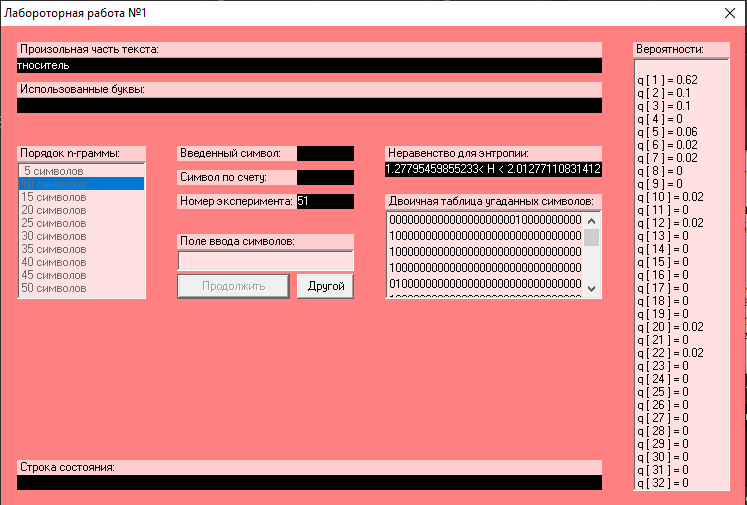
Ентропія біграм з пробілами та перетинами: 3,97272

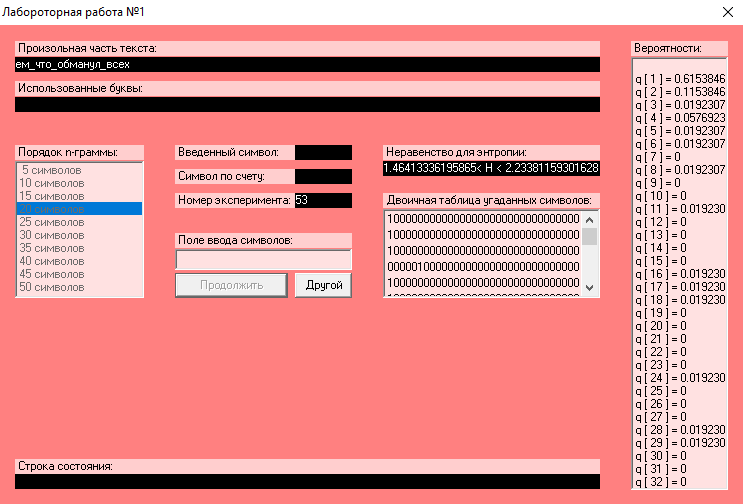
Надлишковість для джерела біграм з пробілами та перетинами: 0,205457

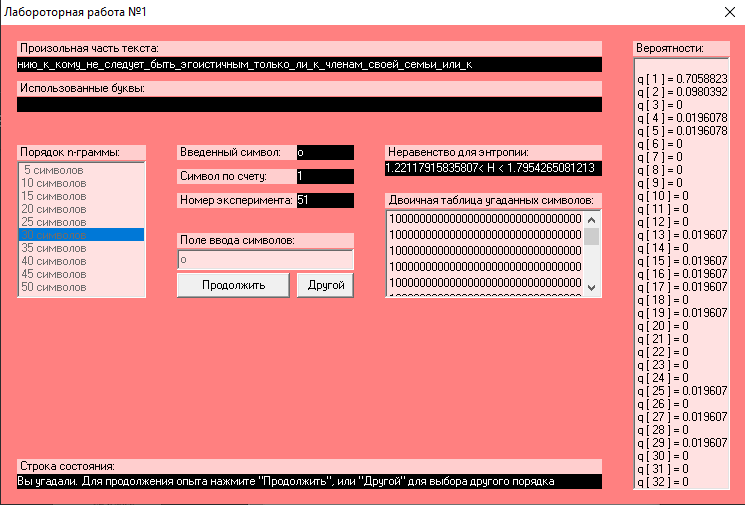
Ентропія біграм з пробілами та без перетинів: 3,95689

Надлишковість для джерела біграм з пробілами та без перетинів: 0,208623

Завдання 2:







Обрахуємо надлишковість для H(10), H(20), H(30):

R = 1 -

H0 = log2m = log232 = 5

H(10) ≈ 1.64536

R(10)≈0.670928

H(20) ≈ 1.89897

R(20)≈0.620206

H(30) ≈ 1.508305

R(30)≈0,698339

**Труднощі, що виникли під час виконання практикуму**

В ході роботи зтикнувся з труднощами кодування вводу-виводу. Рішенням цієї проблеми стало використання строк розширених (sizeof(wchar\_t) = 4 байта у компіляторі GNU) символів та встановлення локалі ru\_RU.UTF8 для всіх потокових об’єктів.