Міністерство освіти і науки України
Національний технічний університет України
“Київський політехнічний інститут ім. Ігоря Сікорського”
Фізико-технічний інститут

Лабораторна робота № 1

«Експериментальна оцінка ентропії на символ джерела відкритого тексту»

Виконали:
Студенти 3 курсу:
Групи ФБ-04
Осіпчук Антон
Подима Катерина

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 1H та 2H за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення 1H та 2H на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення 1H та 2H 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

Для виконання був взятий текст із роману "Злочин і кара" Федора Достоєвського. Текст був переведений у нижній регістр, "ё" та "ъ" були видалені.
Потім була порахована кількість монограм та біграм, визначені частоти та ентропії.
Всі дані записані у excel файл.

Результати

Текст з пробілами:

монограми:

Ентропія : 4.372861176229328

Надлішковість: 0.1254277647541343

біграми з перетином:

Ентропія: 1.7725260764486034e-05

Надлищковість: 0.9979547691686357

біграми без перетину:

Ентропія: 0.010001063419265328

Надлішковість: 0.9979997873161469

Текст без пробілів:

монограми:

Ентропія для монограм: 4.45631984231334

Надлішковість: 0.10049591031136418

біграми з перетином:

Ентропія: 0.010372789798715037

Надлішковість: 0.9979062618538213

біграми без перетину:

Ентропія: 0.010113284422322623

Надлішковість: 0.9979586427770092

Найчастіша поява монограм

монограми з пробілом

Буква	Частота

о	0.11575178997613365
е	0.07159904534606205
и	0.05489260143198091
н	0.05369928400954654
а	0.0513126491646778
л	0.0441527446300716
р	0.041766109785202864
к	0.0405727923627685
т	0.03699284009546539
в	0.034606205250596656
с	0.03341288782816229
м	0.029832935560859187
у	0.026252983293556086
д	0.025059665871121718
й	0.021479713603818614
п	0.017899761336515514
я	0.017899761336515514
г	0.013126491646778043
ч	0.013126491646778043
х	0.011933174224343675
ы	0.011933174224343675
з	0.011933174224343675
б	0.01282051282051282
ь	0.01282051282051282
ж	0.01282051282051282
ц	0.008547008547008548
ю	0.008547008547008548
щ	0.005698005698005698
ш	0.004273504273504274
э	0.002849002849002849
ф	0.001424501424501424

Монограми без пробілу

Буква	Частота
о	0.11575178997613365
е	0.07159904534606205
и	0.05489260143198091
н	0.05369928400954654
а	0.0513126491646778
р	0.041766109785202864
к	0.0405727923627685
т	0.03699284009546539
в	0.034606205250596656
с	0.03341288782816229
м	0.029832935560859187
у	0.026252983293556086
д	0.025059665871121718
й	0.021479713603818614
п	0.017899761336515514
я	0.017899761336515514
г	0.013126491646778043
ч	0.013126491646778043
х	0.011933174224343675
ы	0.011933174224343675
з	0.011933174224343675
б	0.010739856801909307
ь	0.010739856801909307
ж	0.010739856801909307
ц	0.007159904534606206
ю	0.007159904534606206
щ	0.00477326968973747
ш	0.003579952267303103
э	0.002386634844868735
ф	0.0011933174224343676

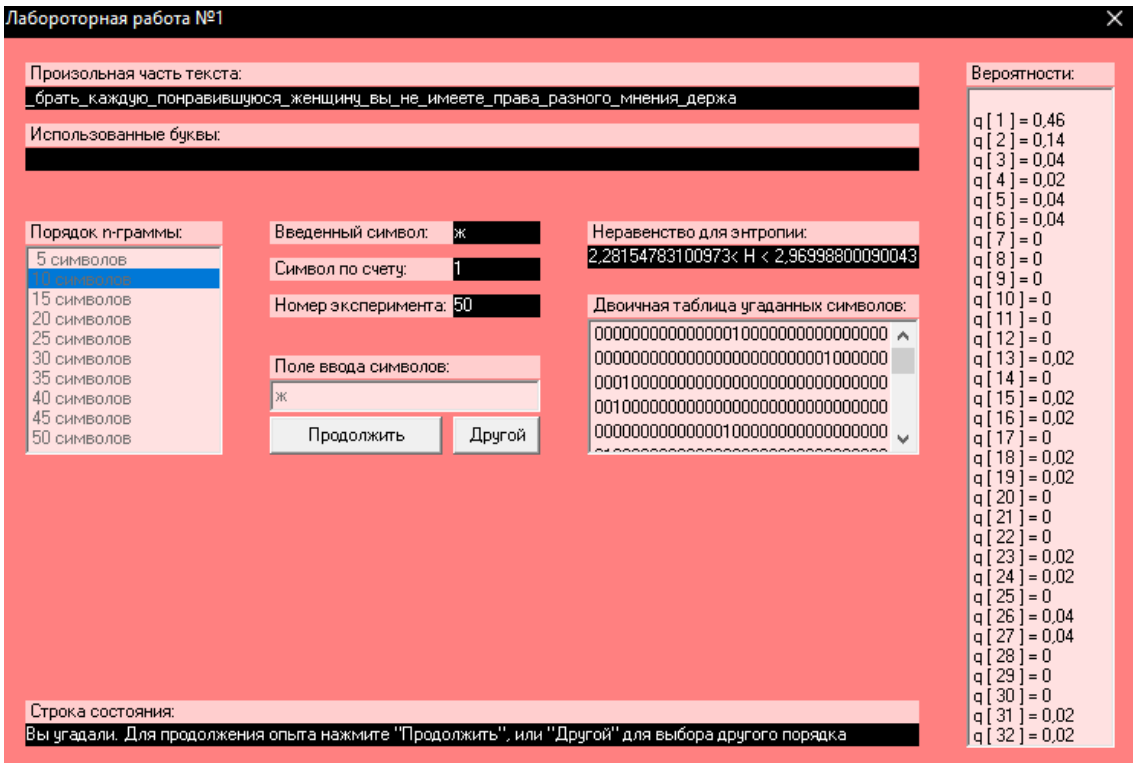
Біграми

Біграми без перетину без пробілу	Частота
го	0.008353221957040573
ст	0.008353221957040573
на	0.008353221957040573
мо	0.008353221957040573
пр	0.007159904534606206
ни	0.007159904534606206
от	0.007159904534606206
од	0.007159904534606206
но	0.0059665871121718375

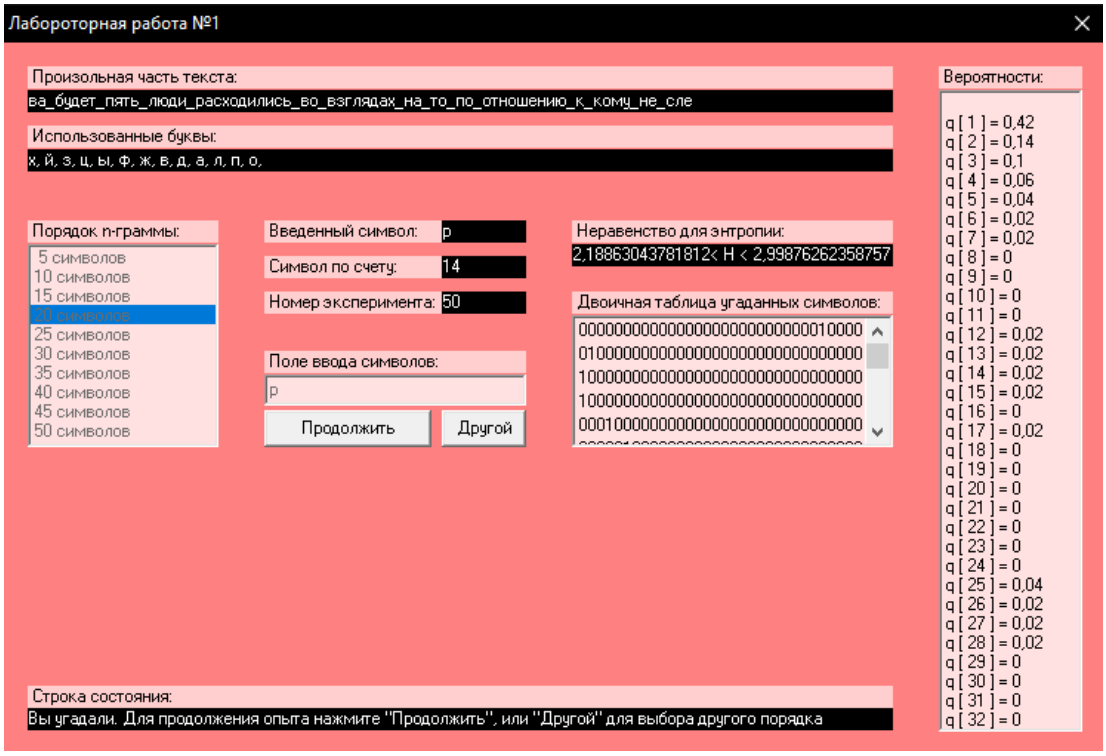
Біграми без перетину з пробілом	Частота
о	0.01207692009459931
е	0.009545225448765168
и	0.00879696041561664
а	0.008513937890125228
в	0.008358857054239522
п	0.00815466728699001
н	0.008068080486953825
с	0.008064203466056682
то	0.007265537161245299

Біграми з перетином без проб	Частота
од	0.014319809069212411
на	0.014319809069212411
мо	0.013126491646778043
ст	0.011933174224343675
ст	0.0046588867780664005
ка	0.011933174224343675
но	0.011933174224343675
хо	0.010739856801909307
го	0.010739856801909307

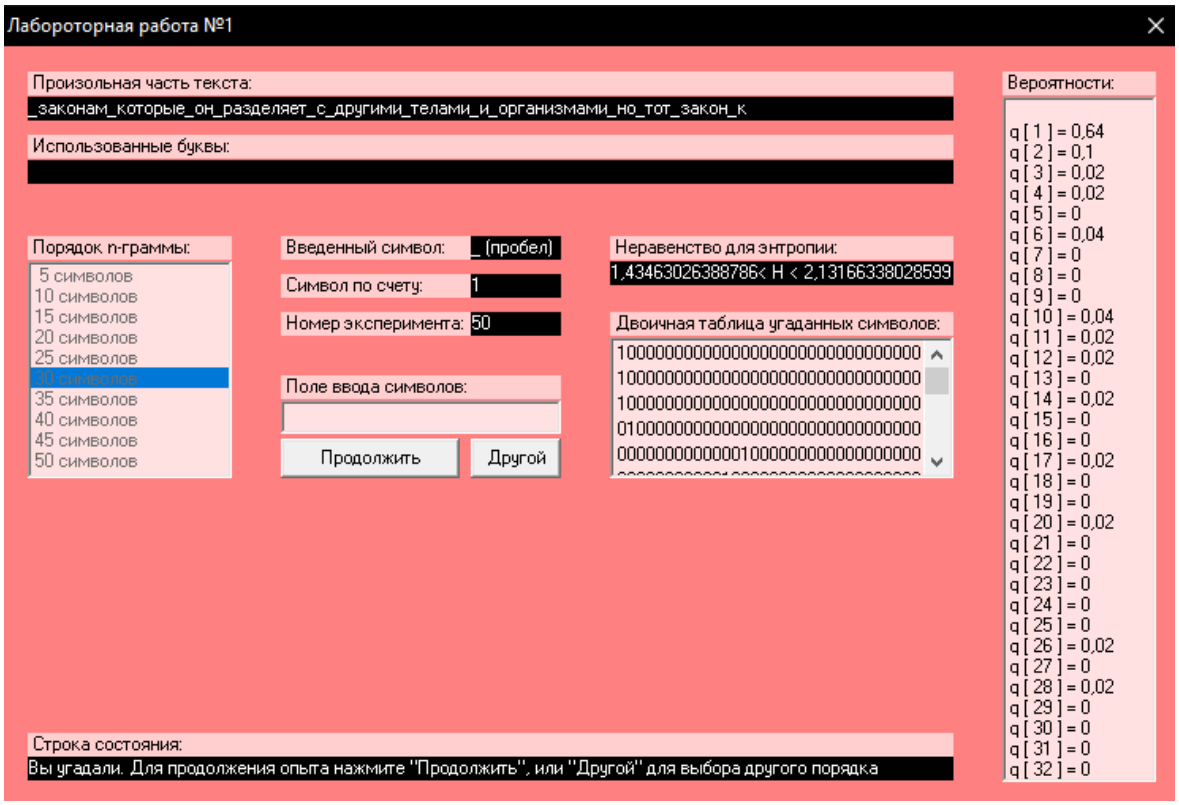
Біграми з перетином з проб	Частота
о	0.024299874642990993
е	0.019095620258726527
и	0.017577120407345662
а	0.017202341720621874
в	0.016691867302498095
п	0.01628865712919526
н	0.016147792036599076
с	0.0160831750216467
то	0.014629292185218211



R = 0,474846416808984



R = 0,481260693859431



R = 0,643370635582615

Висновок

Були засвоєні поняття ентропії на символ джерела та його надлишковості, вивчені та порівняні різні моделі джерела відкритого тексту для наближеного визначення ентропії, отримані практичні навички щодо оцінки ентропії на символ джерела. Після аналізу тексту визначено, що найчастіші у використанні букви російського алфавіту «о», «е», «а».