КРИПТОГРАФІЯ

КОМП’ЮТЕРНИЙ ПРАКТИКУМ №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконали:

ФБ-04 Ковальчук Єгор

ФБ-04 Омелянович Олександр

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 виконання роботи

Був взятий для виконання текст перших двох творів Анджея Сапковського із серії "Відьмак", спершу було створено файл із оригінальним текстом, потім було перетворено його у текст тільки із маленькими буквами, та вирізано букви "ё" та "ъ", також було прибрано всі множинні пробіли, цей текст занесено у файл "clear_text.txt".

Після цього була створена ще одна копія цього тексту, але без пробілів взагалі і записано у файл "withoutspaces.txt".

Було пораховано кількість, частоту та ентропію для кожної монограми для тексту з пробілами/без пробілів.

Було пораховано кількість, частоту та ентропію для кожної біграми з перетинами/без перетинів для тексту з пробілами/без пробілів.

Всі ці дані було занесено у excel файл з назвою "result.xlsx", та уже в ньому було пораховано загальну ентропію та надлишковість.

Результати

Монограми з пробілами

Ентропія = 4,406383

R = 0,118723

Монограми без пробілів

Ентропія = 4,491052

R = 0,093485

Біграми з перетинами і пробілами

Ентропія = 4,011989

R = 0,197602

Біграми з перетинами без пробілів

Ентропія = 4,182021

R = 0,155863

Біграми без перетин з пробілами

Ентропія = 4,012104

R = 0,197579

Біграми без перетин без пробілів

Ентропія = 4,18171

R = 0,155926

Найчастіші монограми та біграми

Монограми з пробілами	
" "	0,16019534
о	0,08862054
е	0,07026795
а	0,06917973
н	0,05388906
и	0,05247785
т	0,05102193
л	0,04500444
с	0,03937553
р	0,03915292
в	0,03554243
к	0,03149764
у	0,02627222
д	0,02599694
м	0,02527046
п	0,0229221
ь	0,01915956
я	0,01874514
ы	0,01688872
г	0,01596448
з	0,01499353
б	0,01463775
ч	0,0126402
й	0,0093517
ж	0,00868486

Монограми без пробілів	
о	0,10552518
е	0,08367178
а	0,08237598
н	0,06416856
и	0,06248817
т	0,06075452
л	0,05358918
с	0,04688654
р	0,04662146
в	0,04232226
к	0,03750592
у	0,03128373
д	0,03095593
м	0,03009088
п	0,02729457
ь	0,0228143
я	0,02232084
ы	0,02011029
г	0,01900975
з	0,01785359
б	0,01742994
ч	0,01505136
й	0,01113557
ж	0,01034152
ш	0,00970368

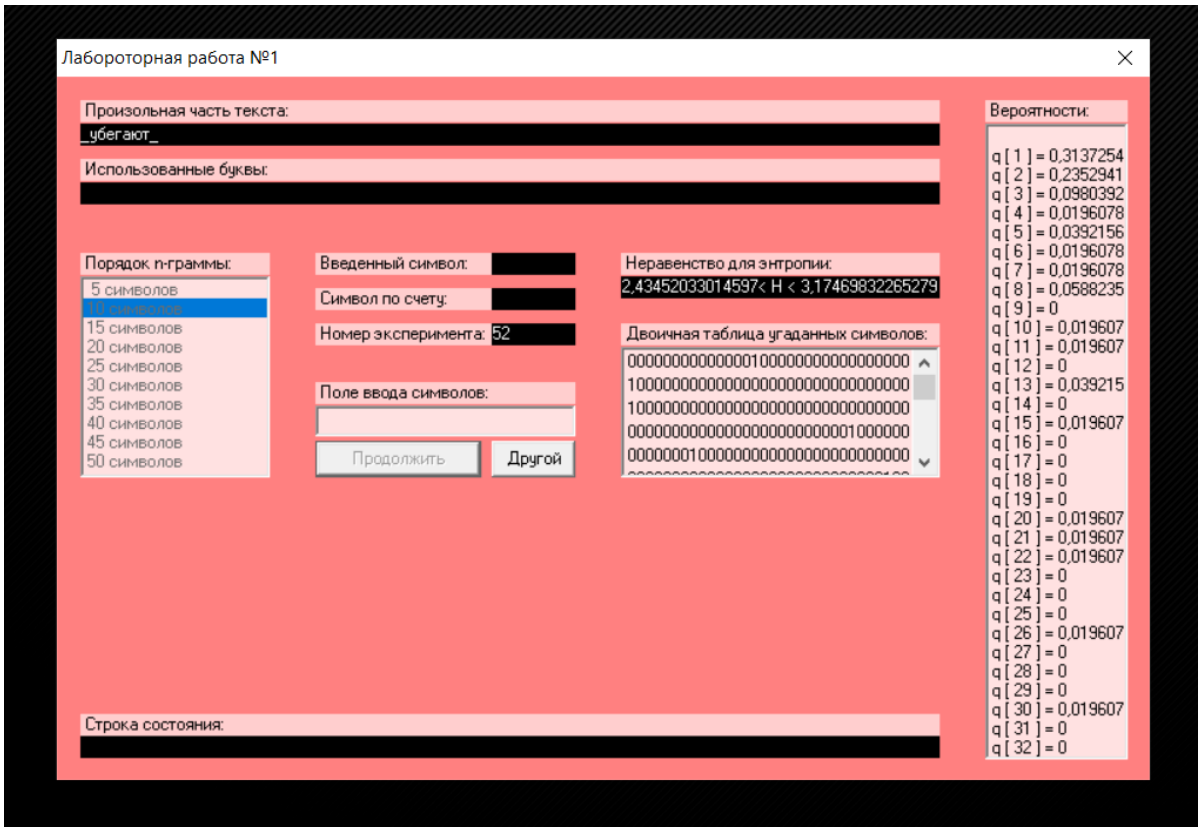
ш	0,0081492
х	0,00719216
ю	0,00585649
щ	0,00348229
э	0,00334713
ц	0,00285421
ф	0,00136549

х	0,00856409
ю	0,00697363
щ	0,00414655
э	0,00398561
ц	0,00339866
ф	0,00162596

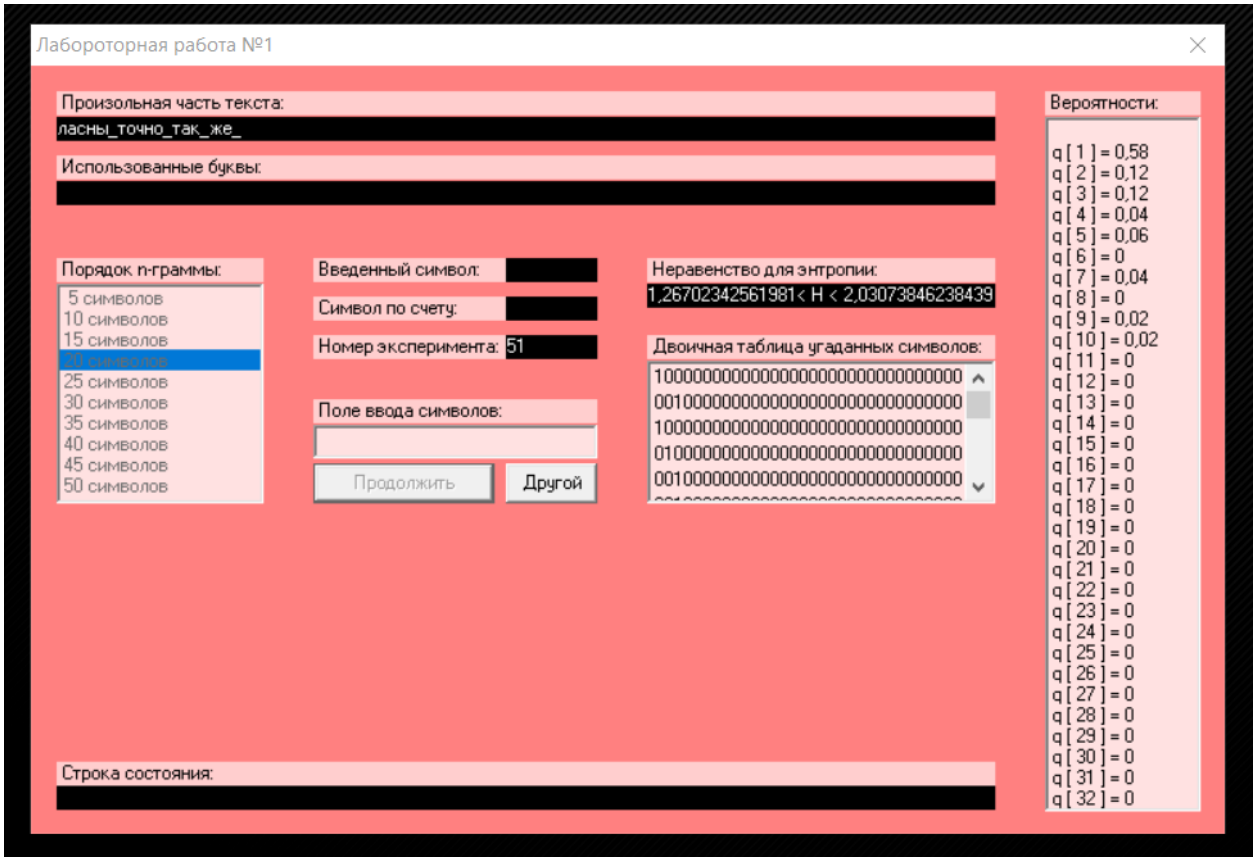
Біграми з перетинами і пробілами		Біграми з перетинами без пробілів	
о	0,019419367	то	0,014330371
а	0,017497307	на	0,01189325
е	0,01692685	не	0,011394936
н	0,0161139	ал	0,011263551
п	0,015901222	по	0,011106127
и	0,015724321	ст	0,011069434
в	0,013965248	но	0,010674097
с	0,013435538	ко	0,010297698
я	0,011784793	ра	0,010032562
то	0,011633731	ен	0,00982069

Біграми без перетин з пробілами		Біграми без перетин без пробілів	
о	0,019327512	то	0,014425353
а	0,017528726	на	0,012113036
е	0,016837038	не	0,011412478
п	0,01596249	по	0,01121367
и	0,015910812	ал	0,011114267
н	0,01582932	ст	0,011052731
в	0,013829785	но	0,010510745
с	0,013386548	ко	0,010162833
я	0,011726895	ра	0,01001136
то	0,011707018	ен	0,009859888

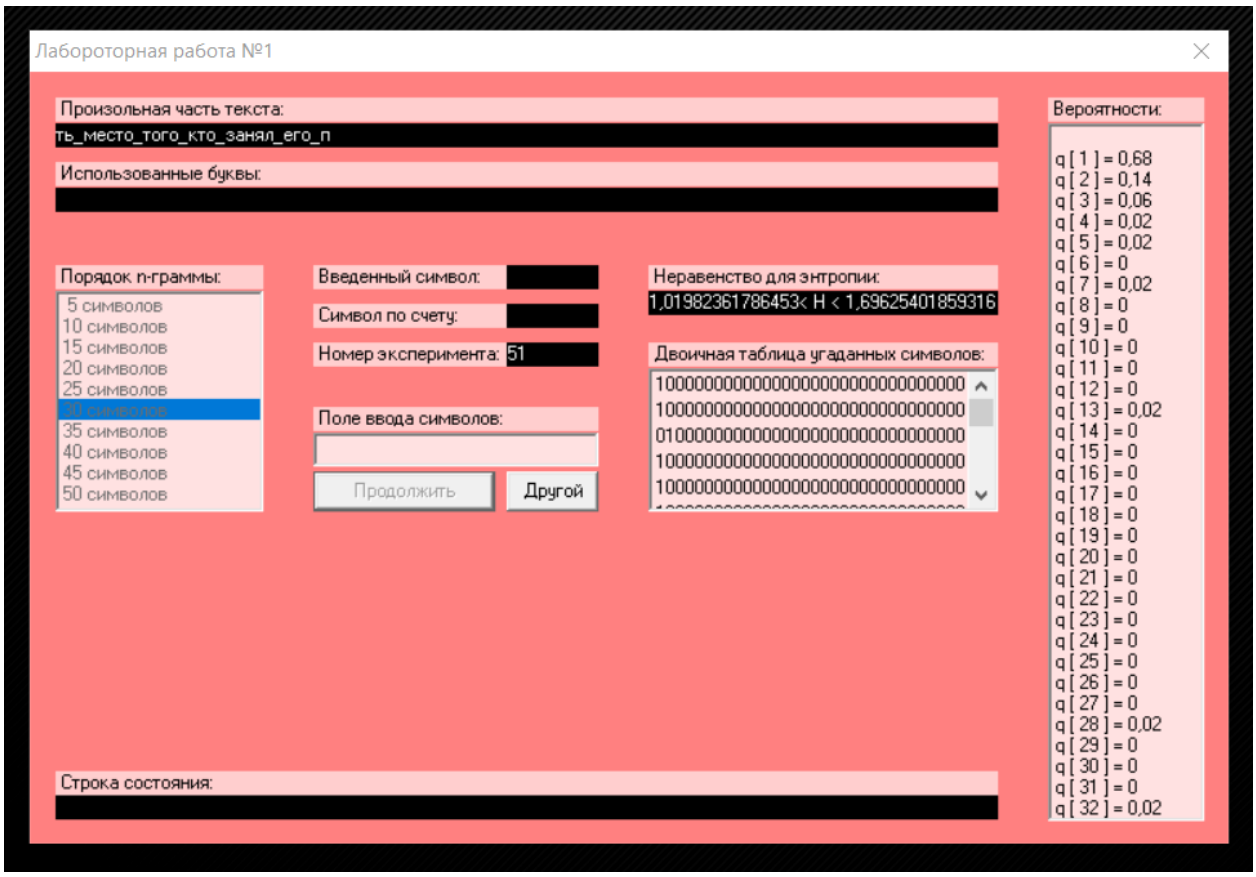
Програма CoolPinkProgram



R = 0,439078134720124



R = 0.67022381119958



R = 0.728392236354231

Висновки

Нами було вивчено поняття ентропії та надлишковості, набуто практичних навичок щодо опрацювання текстів для оцінки ентропії та надлишковості, порівняли різні варіації тексту(з пробілами та без пробілів).

Експериментально було визначено, що в російському тексті найбільш частим символом є “пробіл” у тексті з пробілами, та “о” у тексті без пробілів. Найбільш часто зустрічаються із біграм такі: “о ”, “а ”, “е ” для тексту з пробілами, та “то”, “на”, “не” для тексту без пробілів незалежно від перетинів.