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

КРИПТОГРАФІЯ КОМП’ЮТЕРНИЙ ПРАКТИКУМ №1
Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконали: Студент групи ФБ-05 Даниленко Данило,
Студентка ФБ-05 Мірошніченко Ілона

Київ – 2022

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

Перед виконанням роботи були розглянуті теоретичні відомості в методичних вказівках. В якості експериментального тексту була взята книга «Good Omens» в перекладі на російську мову. Оригінал тексту можна знайти у файлі badtxt.txt. Відредагований текст з пробілами міститься у файлі spaces.txt, а без пробілів у nospaces.txt. В ході виконання роботи було прийнято рішення пробіли замінити на «_» для кращого сприйняття. Усі таблиці також наведені у файлах з відповідними назвами. Єдина відмінність - таблиці біграм були виконанні у двох варіантах(таблиці у вигляді матриці і звичайні)

Monograms with space

h1 with space - 4.401616273407084
r1 with space - 0.13481112083654945

и - 0.06327	у - 0.029778	щ - 0.003901
_ - 0.156775	х - 0.009649	ш - 0.008112
в - 0.035572	ы - 0.01596	б - 0.015512
о - 0.086727	а - 0.074319	г - 0.013423
с - 0.042807	я - 0.018134	ж - 0.009027
ь - 0.015801	п - 0.022487	ф - 0.000767
м - 0.02446	ч - 0.011965	э - 0.002737
е - 0.066044	л - 0.036091	ц - 0.003482
р - 0.04054	н - 0.055251	ё - 0.000143
к - 0.031018	й - 0.009144	ъ - 0.000152
з - 0.014949	т - 0.052805	
д - 0.024208	ю - 0.00499	

Monograms without space

h1 without space - 4.4769563889834885
r1 without space - 0.11248877802734647

в - 0.042185	у - 0.035314	ю - 0.005917
о - 0.102851	х - 0.011443	щ - 0.004627
с - 0.050766	ы - 0.018928	ш - 0.009621
ь - 0.018739	а - 0.088137	б - 0.018396
м - 0.029008	я - 0.021505	г - 0.015919
е - 0.078324	п - 0.026668	ж - 0.010705
р - 0.048078	ч - 0.014189	ф - 0.00091
к - 0.036785	л - 0.042801	э - 0.003246
и - 0.075032	н - 0.065523	ц - 0.004129
з - 0.017728	й - 0.010844	ё - 0.00017
д - 0.028709	т - 0.062623	ъ - 0.00018

Bigrams with space

h2 with space - 4.00310786467689
r2 with space - 0.21314258411506126

и_ - 0.021861	в_ - 0.006095	кр - 0.002883
во - 0.005442	зд - 0.000869	ик - 0.003317
сь - 0.002842	ух - 0.000468	ив - 0.002595
ме - 0.003845	е_ - 0.017426	ая - 0.002925
рк - 0.000185	вы - 0.002619	_и - 0.011286

Bigrams without space

h2 without space - 4.161139326806442

r2 without space - 0.17509630922025265

во - 0.007272	зд - 0.001107	ки - 0.004951
сь - 0.003263	ух - 0.000675	ва - 0.007243
ме - 0.004754	ев - 0.003706	яи - 0.001078
рк - 0.000266	ык - 0.000637	по - 0.010206
ив - 0.005622	ри - 0.0056	

Bigrams with spaces with intersection

h2 with space with intersection - 4.002665513513063

r2 with space with intersection - 0.21322953338184325

_и - 0.011321	ки - 0.003788	ри - 0.004719
_в - 0.015072	оз - 0.001494	ва - 0.005898
ос - 0.005984	ду - 0.001639	я_ - 0.011541
ьм - 0.000163	хе - 0.00022	и_ - 0.021899
ер - 0.005741	ык - 0.000234	по - 0.008672

Bigrams without space with intersection

h2 without space without intersection - 4.162934395360695

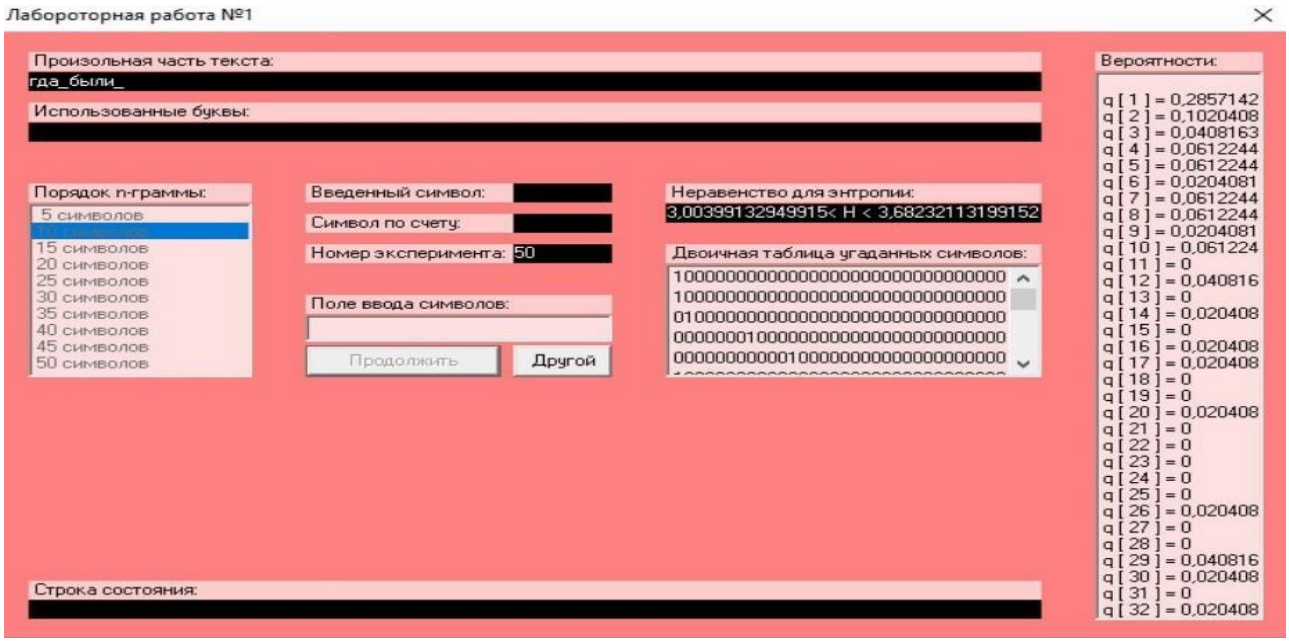
r2 without space without intersection - 0.1747404550756756

ив - 0.005686	вв - 0.000588	кр - 0.003534
ос - 0.00934	оз - 0.002413	ик - 0.005438
ьм - 0.000596	ду - 0.002011	ая - 0.003512
ер - 0.007533	хе - 0.000301	ип - 0.002895
ки - 0.004803	вы - 0.003028	од - 0.005559

Результати експериментів у CoolPinkProgram

n=10:

0.4095305610604002 < R < 0.2761969478903318

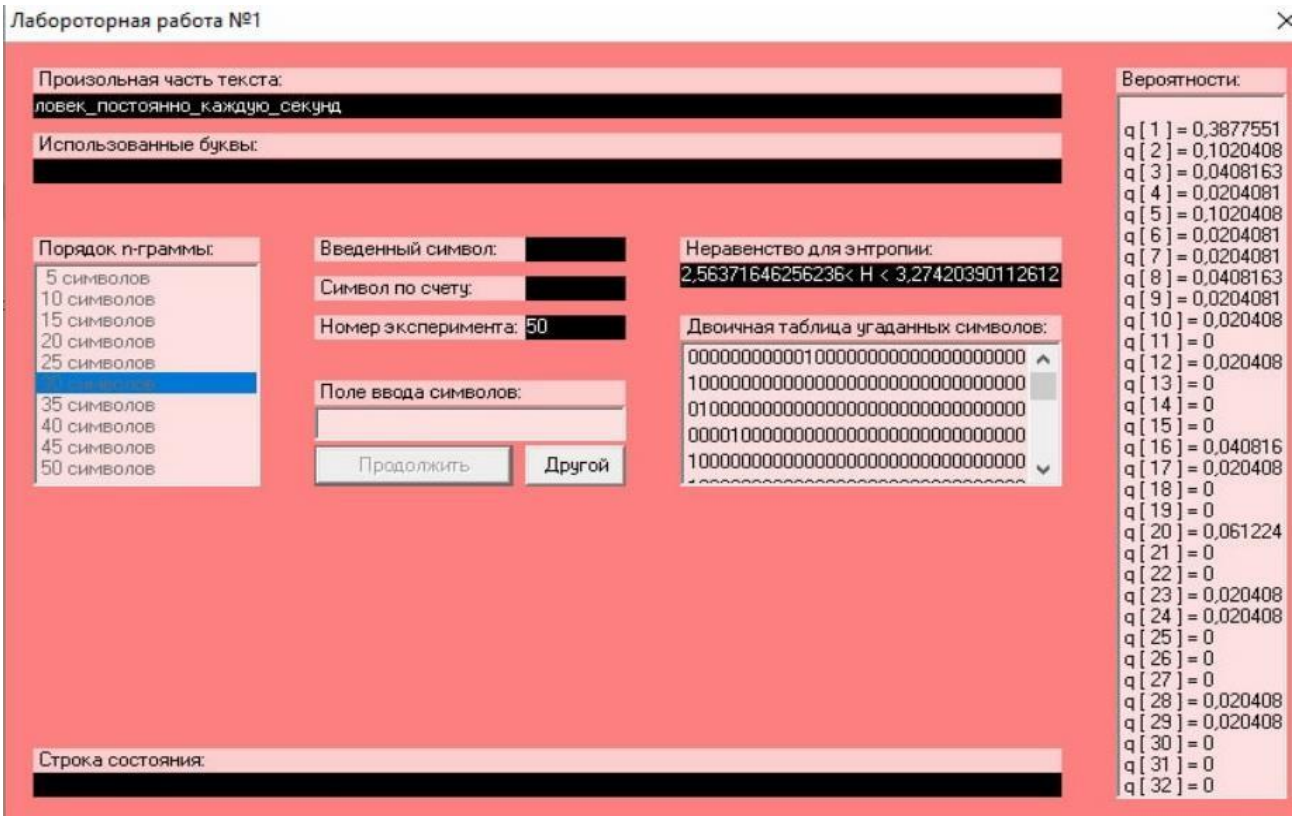


n=20:

0.7572249602844374 < R < 0.6279753922370384

n=30:

0.49607170753658447 < R < 0.3564171369342296



Висновки:

Під час виконання даної лабораторної роботи, ми здобули навички з аналізу тексту. Використовуючи regex нам вдалося в досить простій формі отримати працююче рішення. Крім того, нами було засвоєно визначення ентропії та надлишковості, які ми інтегрували в наш розв’язок. В результаті ми перевірили статистику тексту за допомогою CoolPinkProgram. На нашу думку, дані навички стануть у нагоді у подальшому розвитку.