Міністерство освіти і науки України
Національний технічний університет України
"Київський політехнічний інститут імені Ігоря Сікорського"
Фізико-технічний інститут

Криптографія
Комп’ютерний практикум №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконали
студенти групи ФБ-13
Сварник Назар та Шматко Андрій

Київ - 2023

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

1. Написати програми для підрахунку частот букв і частот біграм в тексті, а також підрахунку H_1 та H_2 за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення H_1 та H_2 на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення H_1 та H_2 на тому ж тексті, в якому вилучено всі пробіли.

Букви з пробілами:

H_1 = 4.376985544110652;

R = 0.124602891;

	0.1644411747153852
о	0.09524140285725187
е	0.06955783327838493
а	0.06563676723366246
н	0.05448403447172733
т	0.052636985396768254
и	0.051725200577985485
л	0.045126500360507193
с	0.041996887017925574
р	0.03723936601605798
в	0.03446487916405377
к	0.03127656722798827
д	0.028515776714463627
у	0.025531931545700277
м	0.025144520828706742
п	0.022873863570772406
ь	0.0168660625276984
я	0.016299621100579063
ы	0.01626929349394573
г	0.014884006687726417
з	0.014541598225737182
б	0.014349849487023208
ч	0.01303891423255013
ж	0.0091501324142438
й	0.009125674666958855

ш	0.006962631497078277
х	0.006827624732065379
ю	0.004685126069904155
э	0.004401416201398787
щ	0.0033008175735762414
ц	0.002549475576982716
ф	0.0008540645351902959

Букви без пробілів:

H_1 = 4.466659028062471;

R = 0.106668194;

о	0.11398527545299994
е	0.08324708108335949
а	0.07855433423410343
н	0.06520670098023146
т	0.06299614557818668
и	0.06190491801743154
л	0.05400756834222395
с	0.05026203511598391
р	0.044568215772687464
в	0.04124769928953124
к	0.03743191536195503
д	0.03412779070911058
у	0.03055671339118869
м	0.03009305876236998
п	0.02737552746568253
ь	0.020185368183925703
я	0.019507448916031678
ы	0.019471152720341323
г	0.017813236168807752
з	0.017403440411013438
б	0.01717395478664862
ч	0.015605022456807527
ж	0.010950913493286375
й	0.010921642367729637
ш	0.008332904023491834
х	0.008171327410418648
ю	0.005607176811648503
э	0.005267631755190356
щ	0.003950431105137199
ц	0.003051222128034245
ф	0.0010221477044412494

Біграми, що не перетинаються без пробілів:

H_2 = 4.154146482675687;

R = 0.169170703;

то	0.016775867479077
ст	0.012741135532336398
но	0.01184895162536706
на	0.011394663756726505
не	0.011125369401604525
он	0.011120686021515449
ал	0.011045751940090201
по	0.010945059268175027
ко	0.010429887458376459
ло	0.009814022976662716
от	0.009591562422431517
го	0.009123224413523728
ен	0.009090440752900183
ос	0.009085757372811105
ка	0.009015506671474938

Біграми, що не перетинаються з пробілами:

H_2 = 3.9760186154157737;

R = 0.204796277;

о	0.02229179099367033
и	0.01776612500856022
е	0.017102830253284677
а	0.016821076728919845
с	0.016091256835947053
п	0.016069733997280296
н	0.015596231546611621
в	0.014263772170969604
то	0.01360439065908802
о	0.011841474510110842
ь	0.010687067708893824
ст	0.01055401743349932
я	0.010013989845133393
но	0.00987898294804191
т	0.009556140368040542

Біграми, що перетинаються без пробілів:

H_2 = 4.154812266889686;

R = 0.169037547;

то	0.016575672387812424
ст	0.012724758600744892
но	0.012026934150428003
на	0.011212025060796198
он	0.011094940421481285
по	0.011050448258541618
не	0.011005956095601949
ал	0.01096497647184173
ко	0.010219147319405725
от	0.009700462367240655
ло	0.009581036035139443
ос	0.009144310330494811
ов	0.009099818167555144

ка	0.00903776330871824
ен	0.009027225691179898

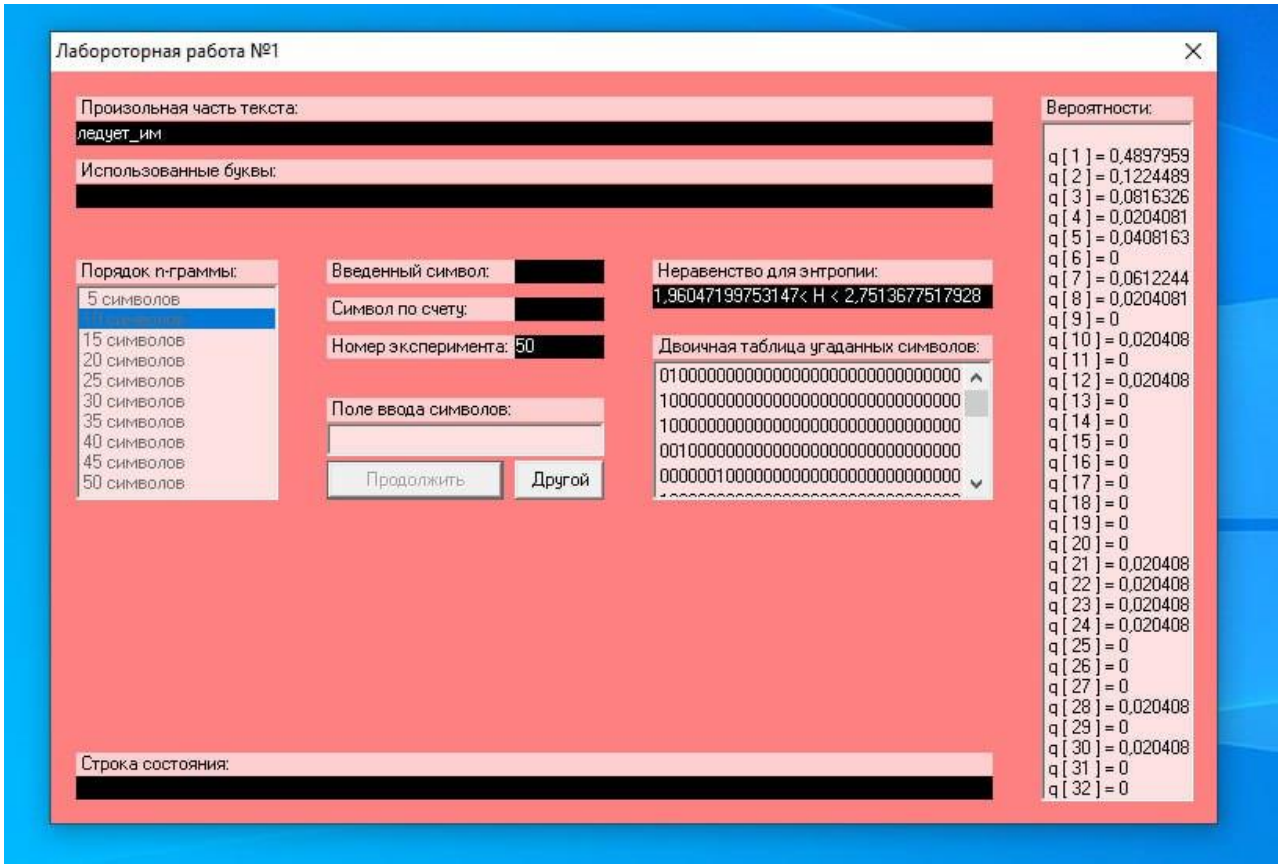
Біграми, що перетинаються з пробілами:

H_2 = 3.9766401969525007;

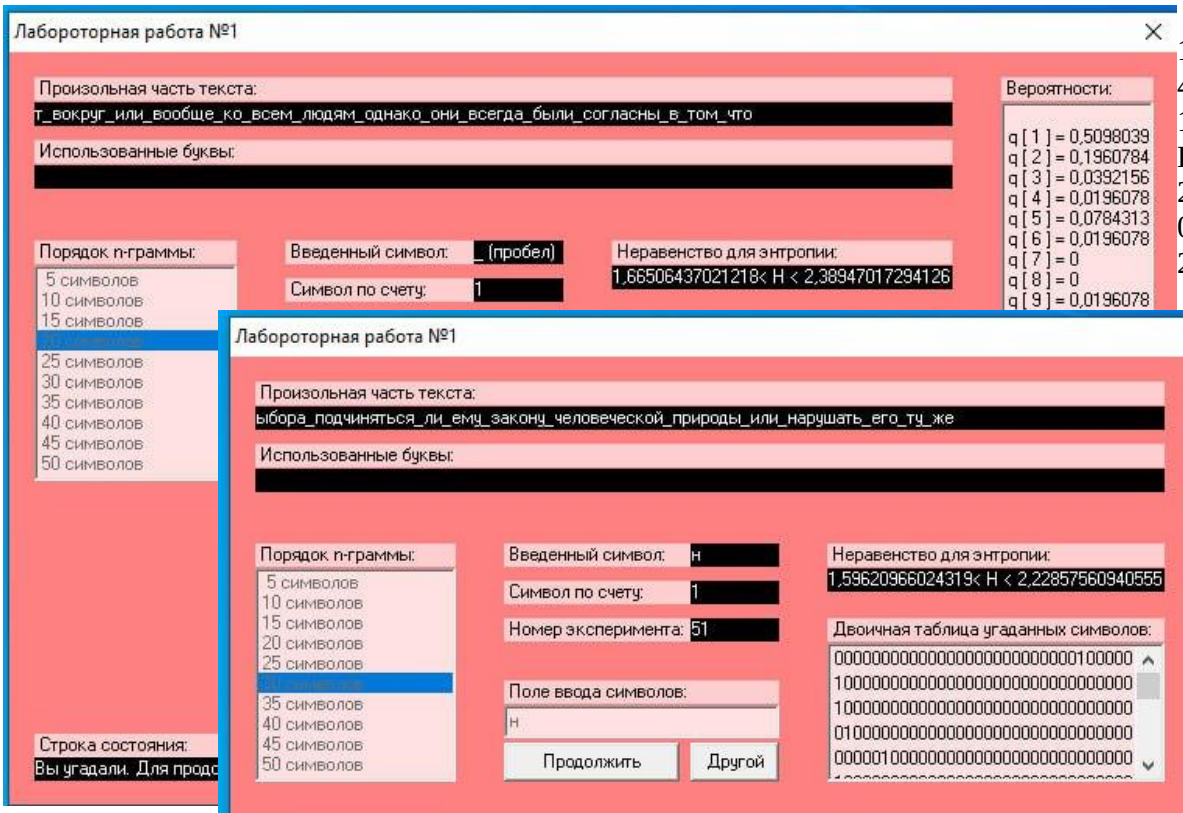
R = 0.204671961;

о	0.022542238570883513
и	0.017721122709529726
е	0.017039240048132894
а	0.01682890321570776
п	0.016125497715644168
с	0.01567645303618772
н	0.01563634229139967
в	0.014256923995030181
то	0.013444925990784312
о	0.011725055519140652
ь	0.010590214934893413
ст	0.01038868290010468
я	0.009893657610769246
но	0.00974299774010194
т	0.009667667804768286

2. За допомогою програми CoolPinkProgram оцінити значення $H^{(10)}$, $H^{(20)}$, $H^{(30)}$.



$1.96047199753147 < H^{(10)} < 2.7513677517928$



$1.66506437021218 < H^{(20)} < 2.38947017294126$

$1.59620966024319 < H^{(30)} < 2.22857560940555$

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

$H_0 = \log_2 32 = 5$

$H^{(10)} = 1 - (1.96047199753147 / 5) = 0.6079056 < R < 1 - (2.7513677517928 / 5) = 0.44972645$

$H^{(20)} = 1 - (1.66506437021218 / 5) = 0.666987126 < R < 1 - (2.38947017294126 / 5) = 0.522105965$

$H^{(30)} = 1 - (1.59620966024319 / 5) = 0.680758068 < R < 1 - (2.22857560940555 / 5) = 0.554284878$

Проблеми, які виникли у ході роботи

Під час прочитання завдання про біграми у методичці не вистачало прикладу для повного розуміння.

Висновки

Під час виконання роботи ми засвоїли поняття ентропії на символ джерела та його надлишковості, набули практичних навичок щодо оцінки ентропії на символ джерела. Також Написали програму для підрахунку частот букв і частот біграм в тексті, а також підрахунку H_1 та H_2 за безпосереднім означенням.