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-11
Комар Анастасія та Сергеєв Максим

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: $H_1 = 4.390654874978746$; $R = 0.12186902500425079$

Літера	Частота
	0.1438891384261639
о	0.09215708848916081
е	0.07197213542468991
и	0.07120028149966677
а	0.06858635602248754
т	0.06000191341939397
н	0.055387005287848005
с	0.04746090867962711
р	0.04010721634773641
л	0.036877104964026096
в	0.03564473315096393
к	0.03182113744675262
п	0.02581981102551172
м	0.023481547664412183
д	0.023001571274061654
з	0.02031305487135498
я	0.01850503569824403
у	0.017316445515382757
г	0.016369465069556038
ь	0.014420371754686662
б	0.012263721081827867
ы	0.011337820680171948
й	0.011141614115197577
ч	0.01050921279007357
ю	0.009468182916407896
ж	0.008949289521434351
х	0.007008303915861436
ш	0.005257038707825723
щ	0.0036598199764227815
ц	0.002437177414516367
э	0.0022166477216526105

ф	0.0014188491268807858

- Без пробілів: H_1 = 4.434379773006463; R = 0.10492449326050612

Літера	Частота
о	0.10764620871617411
е	0.08406870962059698
и	0.08316712787497561
а	0.08011387204736714
т	0.07008661624627577
н	0.06469606656552283
с	0.055437806959226156
р	0.046848157344955406
л	0.04307515138428786
в	0.04163565111816971
к	0.03716941213460843
п	0.030159424654472584
м	0.027428162307443164
д	0.026867514835376097
з	0.023727131360081522
я	0.021615232943342406
у	0.02022687281825741
г	0.019120730508503467
ь	0.016844047192879776
б	0.01432492172717303
ы	0.0132434024482869
й	0.013014218853286512
ч	0.012275527927252202
ю	0.011059529018241877
ж	0.010453423643034239
х	0.008186210723898166
ш	0.006140605082572387
щ	0.004274936974511375
ц	0.002846801184178377
э	0.0025892063997151305
ф	0.0016573193853333863

Примітка: надалі в таблицях буде наведено 10 найбільш вживаних біграм. Щоб ознайомитись з повним набором біграм, перегляньте відповідні csv файли.

Біграми, що перетинаються:

- З пробілами: H_2 = 4.051308552537498; R = 0.21530808592222384

Біграма	Частота
о	0.021290913660723696
п	0.018419153714559062
н	0.017812695092387646
и	0.01617493250310121
ст	0.015164708648521555

ни	0.014357178183705074
с	0.01375396265576987
е	0.01356099854871533
то	0.013251283049157202
в	0.013077777507519924

- Без пробілів: H_2 = 3.912380988415423; R = 0.1822470692080569

Біграма	Частота
ст	0.017925539774126455
ни	0.01702196982038469
то	0.016144919739499077
пр	0.013614166158369103
ос	0.013223945172057147
ра	0.01279205009982838
но	0.012418877603209663
ко	0.01155887601201729
по	0.01091103340367414
ен	0.010545438013585751

Біграми, що не перетинаються:

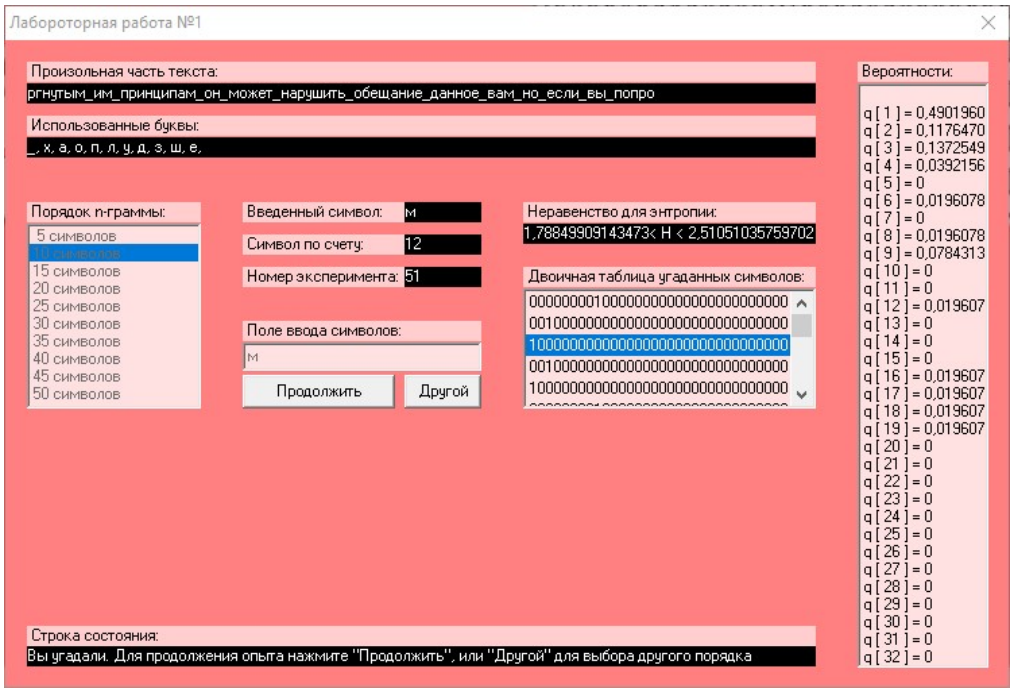
- З пробілами: H_2 = 3.912380988415423; R = 0.21752380231691537

Біграма	Частота
о	0.021477340015899446
п	0.018580434519467007
н	0.01796866565686637
и	0.01631656257872026
ст	0.015297493056259834
ни	0.014482891736540267
с	0.013874394365183482
е	0.013679740636174187
то	0.01336731322238616
в	0.013192288440839988

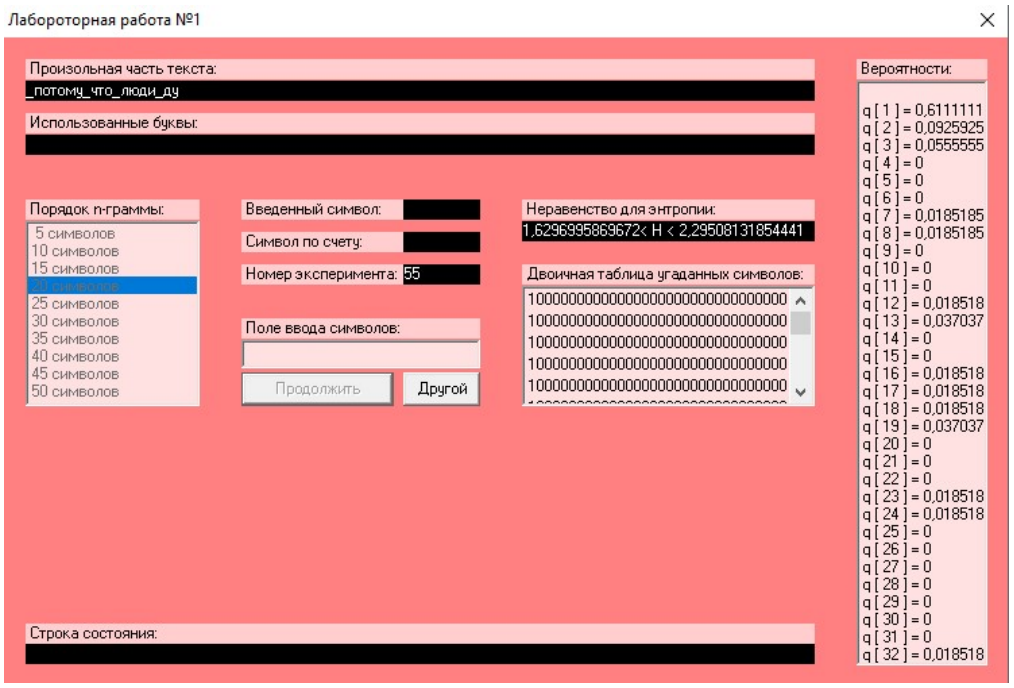
- Без пробілів: H_2 = 4.04913688013438; R = 0.18268541929898174

Біграма	Частота
ст	0.017944030855465507
ни	0.017039528824602454
то	0.01616157402315677
пр	0.01362820984447116
ос	0.013237586325901373
ра	0.012805245732532871
но	0.012431688290017104
ко	0.011570799564625438
по	0.010922288674572685
ен	0.010556316154747593

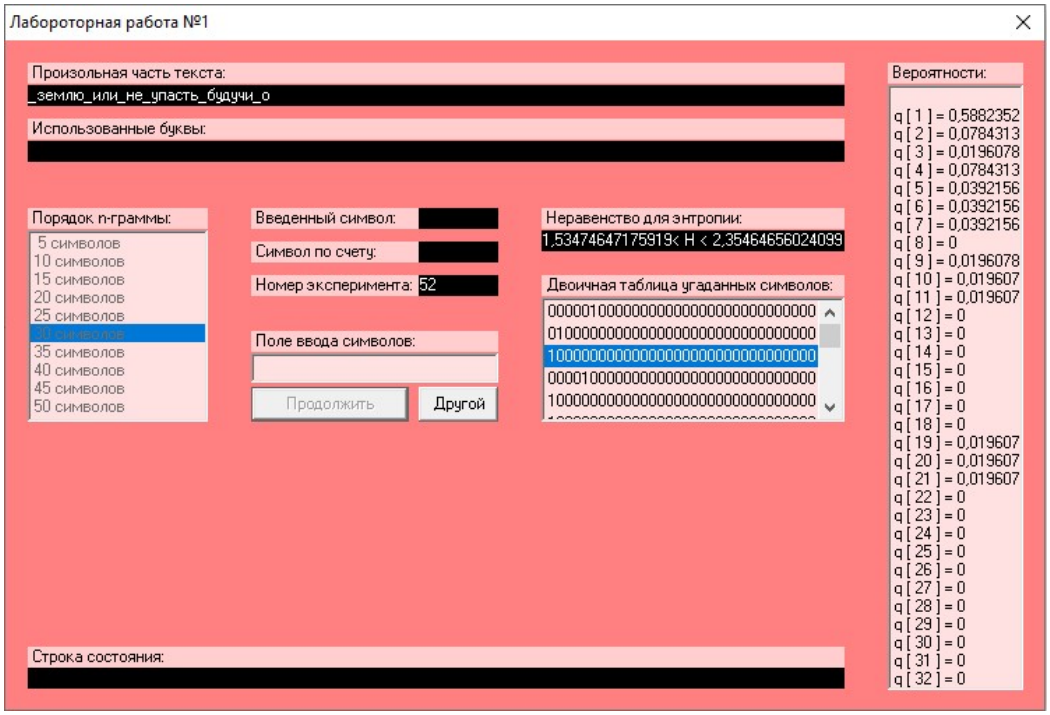
2. За допомогою програми CoolPinkProgram оцінити значення $H^{(10)}$, $H^{(20)}$, $H^{(30)}$.



$1{,}78849909143473 < H^{(10)} < 2{,}51051035759702$



$1{,}6296995869672 < H^{(20)} < 2{,}29508131854441$



$1{,}53474647175919 < H^{(30)} < 2{,}35464656024099$

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

$R = 1 - \frac{H_\infty}{H_0}$; $H_0 = \log_2 32 = 5$

$H^{(10)}$: $1 - \frac{1{,}78849909143473}{5} \approx 0{,}64 < R < 1 - \frac{2{,}51051035759702}{5} \approx 0{,}49$

$H^{(20)}$: $1 - \frac{1{,}6296995869672}{5} \approx 0{,}67 < R < 1 - \frac{2{,}29508131854441}{5} \approx 0{,}54$

$H^{(30)}$: $1 - \frac{1{,}53474647175919}{5} \approx 0{,}69 < R < 1 - \frac{2{,}35464656024099}{5} \approx 0{,}53$

Проблеми, які виникли у ході роботи

1. У методичних вказівках є слова: “прописні літери – замінюватись на відповідні стрічні”, які викликали тимчасові непорозуміння, оскільки в українській мові прописна літера – це літера, яка написана від руки, а такого поняття як “стрічна літера” взагалі не існує. Словом, знадобився час, щоб зрозуміти, що тут говорилось про великі та маленькі літери.

2. Довго думали, чому ентропія біграм завелика, але потім помітили, що пропустили множник у формулі.

Висновки

Під час виконання лабораторної роботи ми ознайомились з поняттями ентропії на символ джерела та його надлишковості, написали скрипт та попрацювали з CoolPinkProgram, які дозволили порівнювати різні моделі джерела відкритого тексту для наближеного визначення ентропії. У підсумку, ми набули практичних навичок оцінки ентропії, що буде корисним у нашій подальшій професійній діяльності.