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:
Курченко Максим
Мартиненко Денис
Група: ФБ-04

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

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 H(10), H(20), H(30).

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

Хід роботи

Спочатку знайшли текст, потім його відфільтрували, залишивши тільки пробіли і всі літири російського алфавіту. Далі за допомогую пайтона написали функції для обрахування монограм і біграм. Всі результати були записані в таблицю excel.

Результати

Монограми з пробілами

Ентропія 4.417161441924015

Надлишковість 0.11656771161519708

Монограми без пробілів

Ентропія 4.4637749113737595

Надлишковість 0.09899111143111283

Біграми без перетину з пробілами

Ентропія 4.044355577082975

Надлишковість 0.19112888458340505

Біграми без перетину без пробілів

Ентропія 4.165380067803057

Надлишковість 0.15922183804666767

Біграми з перетином з пробілами

Ентропія 4.045451730369846

Надлишковість 0.19090965392603076

Біграми з перетином без пробілів

Ентропія 4.164782693547804

Надлишковість 0.15934241749443823

Таблиці

Монограми з пробілами	
	0,143178362
о	0,091777385
а	0,072287329
е	0,069185372
и	0,062813144
н	0,053311697
т	0,049395433
с	0,046645513
л	0,042091551
р	0,041892419
в	0,040736741
к	0,031168917
м	0,026546208
п	0,026000965

д	0,025744938
у	0,023985937
ы	0,016606196
я	0,015654392
г	0,015105593
ь	0,01448212
б	0,014331586
з	0,014121786
ч	0,012240699
й	0,010817141
ж	0,008307839
х	0,008028106
ш	0,007149791
ю	0,004735314
ц	0,004649972
щ	0,00282578
ф	0,002128817
э	0,002052957

Монограми без пробілів	
о	0,107114
а	0,084367
е	0,080747
и	0,073309
н	0,06222
т	0,05765
с	0,05444
л	0,049125
р	0,048893
в	0,047544
к	0,036377
м	0,030982
п	0,030346
д	0,030047
у	0,027994
ы	0,019381
я	0,01827
г	0,01763
ь	0,016902
б	0,016726
з	0,016482
ч	0,014286
й	0,012625
ж	0,009696
х	0,00937
ш	0,008345
ю	0,005527

ц	0,005427
щ	0,003298
ф	0,002485
э	0,002396

Неперехресні біграми з пробілами	
и	0,015848
а	0,015842
п	0,01568
о	0,015443
е	0,014608
с	0,014385
в	0,014002
ст	0,012079
н	0,011762
но	0,009885
то	0,009542
на	0,009462
по	0,009193
и	0,009154
о	0,009013
ко	0,00886
ов	0,008756
я	0,008591
ал	0,008554
ро	0,008547
к	0,008509
й	0,0084
ра	0,008277
ли	0,00771
ен	0,007622
ни	0,007508
ь	0,007456
го	0,007288
в	0,006963
не	0,006915
м	0,006914
ор	0,006793
ва	0,006789
ос	0,006609
ка	0,00657 …….

Неперехресні біграми без пробілів	
ст	0,01429
ов	0,012082
но	0,011721
то	0,011467
на	0,011084
по	0,010785
ко	0,010615
ен	0,010314
ал	0,010241
ро	0,010093
ра	0,009696
ли	0,009645
ос	0,009626
ни	0,00897
го	0,008544
ор	0,008531
не	0,00812
ва	0,008044
ер	0,008037
во	0,008026
ка	0,007741
та	0,007685
ол	0,007656
од	0,007567
ан	0,007555
ло	0,007409
ес	0,007347
ом	0,007314
ат	0,007297
ел	0,007292
пр	0,007156
ре	0,007149
он	0,007058
от	0,006938
те	0,006744
ас	0,006592

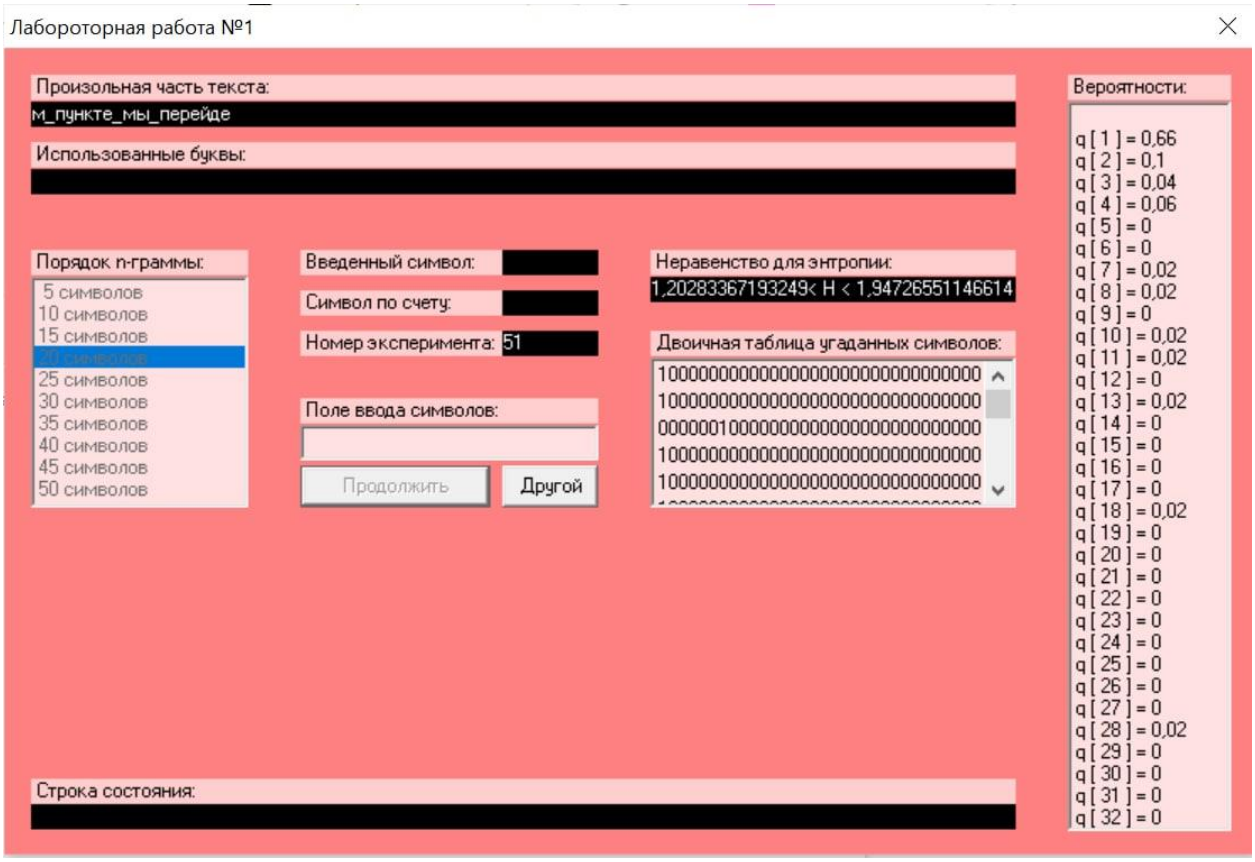
Перехресні біграми з пробілами	
а	0,016047
и	0,01604
п	0,015466
о	0,015243
е	0,014745
с	0,014328
в	0,013776

ст	0,012038
н	0,011735
но	0,00999
то	0,009779
на	0,009357
по	0,009136
и	0,009039
ко	0,009004
о	0,008959
я	0,008767
ал	0,008639
ов	0,008577
к	0,008563
ро	0,008496
ра	0,008437
й	0,008288
ен	0,007769
ли	0,007586
ь	0,007503
ни	0,007496
го	0,00729
в	0,007102

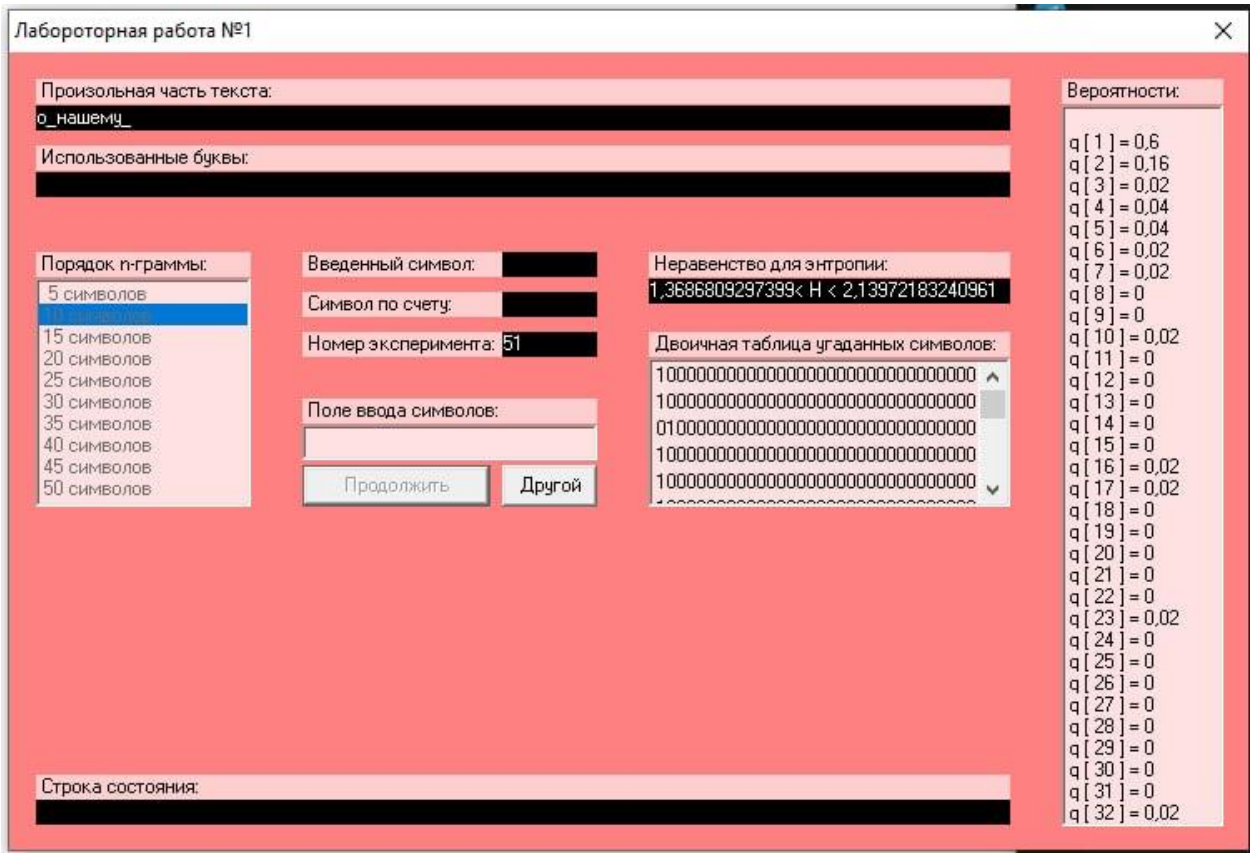
Перехресні біграми без пробілів	
ст	0,014116
ов	0,012035
но	0,011853
то	0,011225
на	0,011183
по	0,01066
ко	0,010522
ен	0,010295
ал	0,010165
ро	0,009944
ос	0,009808
ли	0,009764
ра	0,009706
ни	0,008942
ор	0,008591
го	0,008439
во	0,008179
не	0,008082
ер	0,00806
ва	0,007988
та	0,007866
ка	0,007744

ан	0,007739
ол	0,007609
од	0,007459
ло	0,00744
ат	0,007379
ес	0,007379
ом	0,007304

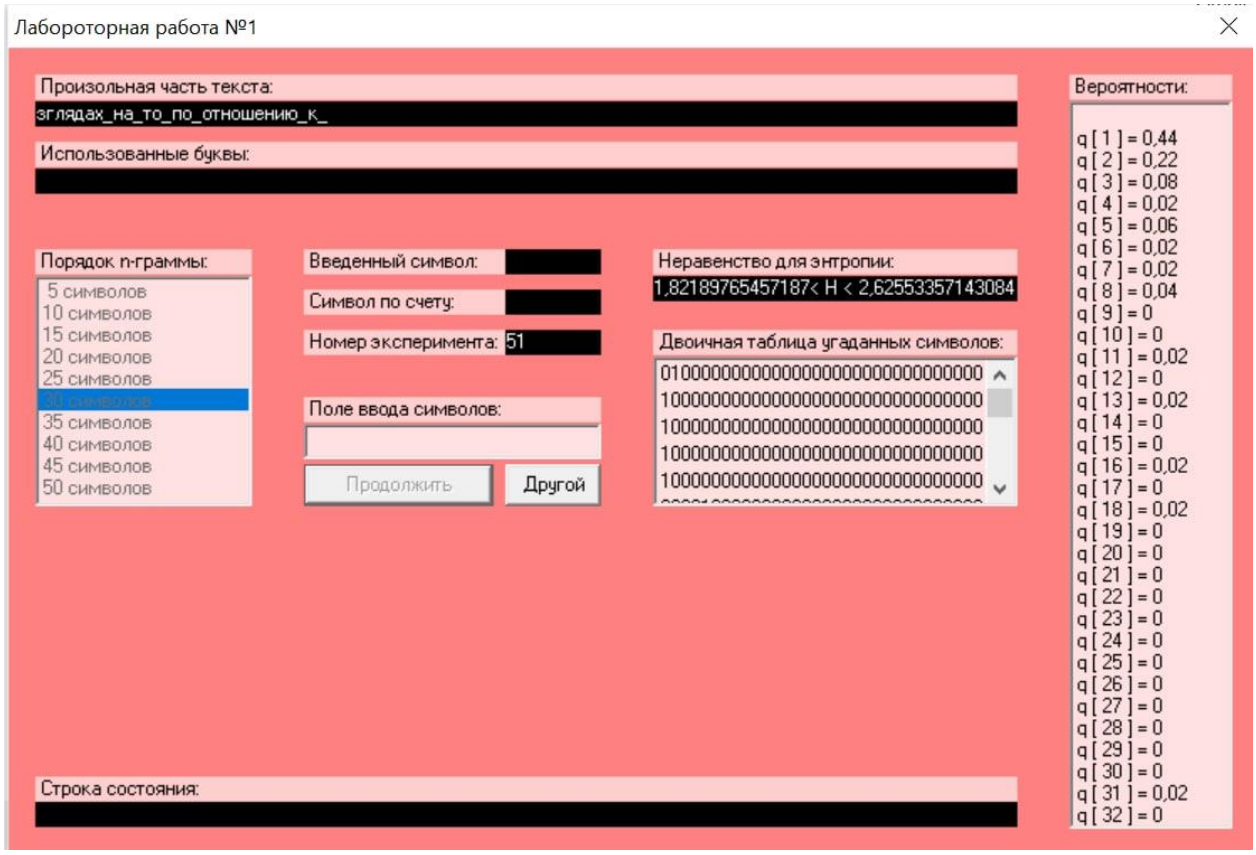
coolpinkprogram



R=0,684991



R=0,64917



R=0,55527

Висновок: В лабораторній роботі ми оцінили частоту появи окремих монограм, біграм в нашому тексті. Найбільші монограми в тексті з пробілами: “ ”, “о”, “а”, “е”, без пробілів “о”, “а”, “е” , “и”, найменші монограми: “ф”, “з”, “ц”, “щ”. З отриманих даних можемо зробити висновок, що найчастіше у нашому тексті зустрічаються букви “о”, “а”, “е”, а найменше “ф”, “з”, “ц”, “щ”.