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

ЛАБОРАТОРНА РОБОТА №1
«Експериментальна оцінка ентропії на символ джерела відкритого тексту»

Виконала:
студентка групи ФБ-04
Андрійчук Анастасія
Перевірив:
Чорний О.

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

Хід роботи

Взяла текст «Преступление и наказание». Очистила текст від непотрібних символів. Далі за формулами порахувала частоти, ентропії, надлишковості. Отримані таблички зберегла через пандас в ексель.

Результат

Монограми з пробілами

ентропія 4.304284399234026

надлишковість 0.13914312015319474

Монограми без пробілів

ентропія 4.399966340633688

надлишковість 0.11187081315110559

Перехресні біграми з пробілами

ентропія 3.9279688039841756

надлишковість 0.21440623920316493

Неперехресні біграми з пробілами

ентропія 3.9273189639991983

надлишковість 0.2145362072001603

Перехресні біграми без пробілів

ентропія 4.101573507324395

надлишковість 0.17210113399723115

Неперехресні біграми без пробілів

ентропія 4.100124677867977

надлишковість 0.17239357889962237

Таблиці

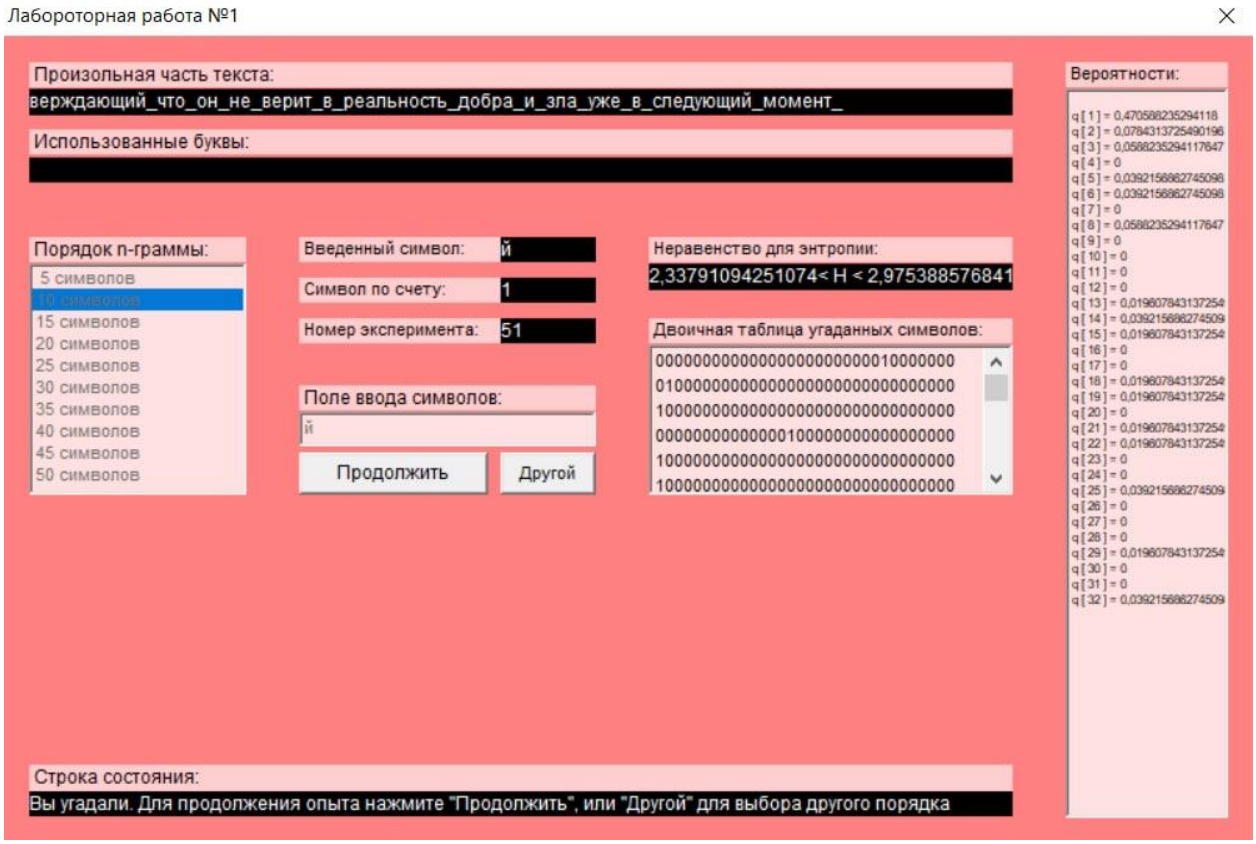
Монограми з пробілами		Монограми без пробілів	
_	0,17253725	**о**	0,116651488
о	0,096524761	**е**	0,08863062
е	0,073338537	**а**	0,080997557
а	0,067022462	**н**	0,066178945
н	0,054760612	**и**	0,065936492
и	0,054559991	**т**	0,065838787
т	0,054479143	**с**	0,053819849
с	0,04453392	**в**	0,047038388
в	0,038922514	**л**	0,046733211
л	0,038669991	**р**	0,042534302
р	0,035195551	**к**	0,033581617
к	0,027787537	**д**	0,03255511
д	0,026938141	**м**	0,031967673
м	0,026452059	**у**	0,030149875
у	0,024947898	**п**	0,027902657
п	0,023088409	**ь**	0,023616899
ь	0,019542105	**я**	0,021723108
я	0,017975063	**ч**	0,018407165
ч	0,015231243	**б**	0,017681011
б	0,014630378	**г**	0,017174392
г	0,014211169	**з**	0,01565574
з	0,012954542	**ж**	0,011600374
ж	0,009598877	**й**	0,01018184
й	0,008425093	**х**	0,008651126
х	0,007158485	**ш**	0,008368867
ш	0,006924926	**ю**	0,005711529
ю	0,004726077	**э**	0,00358614
э	0,002967398	**щ**	0,003040922
щ	0,002516249	**ц**	0,002818974
ц	0,002332596	**ф**	0,001265342
ф	0,001047023		

Перехресні біграми з пробілами		Неперехресні біграми з пробілами		Перехресні біграми без пробілів		біграми без пробілів	
о_	0,0239	**о_**	0,0237571	**то**	0,018448	**то**	0,018436
е_	0,019144	**е_**	0,0191419	**не**	0,01296	**не**	0,013218
и_	0,017574	**и_**	0,017469	**ов**	0,012799	**ов**	0,012846
а_	0,017097	**_в**	0,0170957	**на**	0,012319	**на**	0,012393
_в	0,016902	**а_**	0,0169799	**но**	0,012143	**но**	0,012304
_п	0,016063	**_п**	0,0161954	**ст**	0,011828	**ст**	0,011874
_н	0,016003	**_н**	0,0161854	**по**	0,011074	**по**	0,011184
_с	0,015798	**_с**	0,015914	**ко**	0,010851	**ко**	0,010851
то	0,014845	**то**	0,0149218	**он**	0,010551	**он**	0,010545
ь_	0,01202	**ь_**	0,0120552	**от**	0,00991	**от**	0,009942

CoolPinkProgram

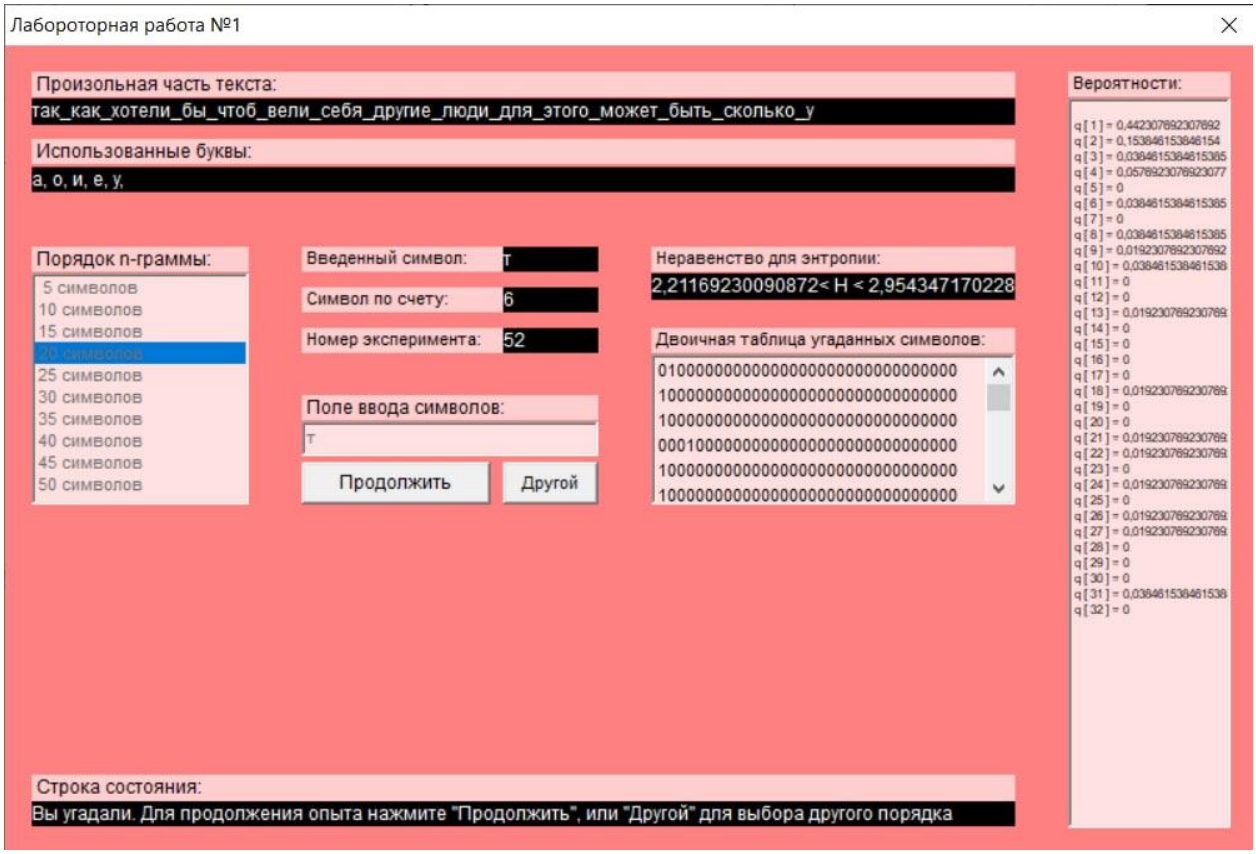
H10 = 2,65664976

R = 0,468670048



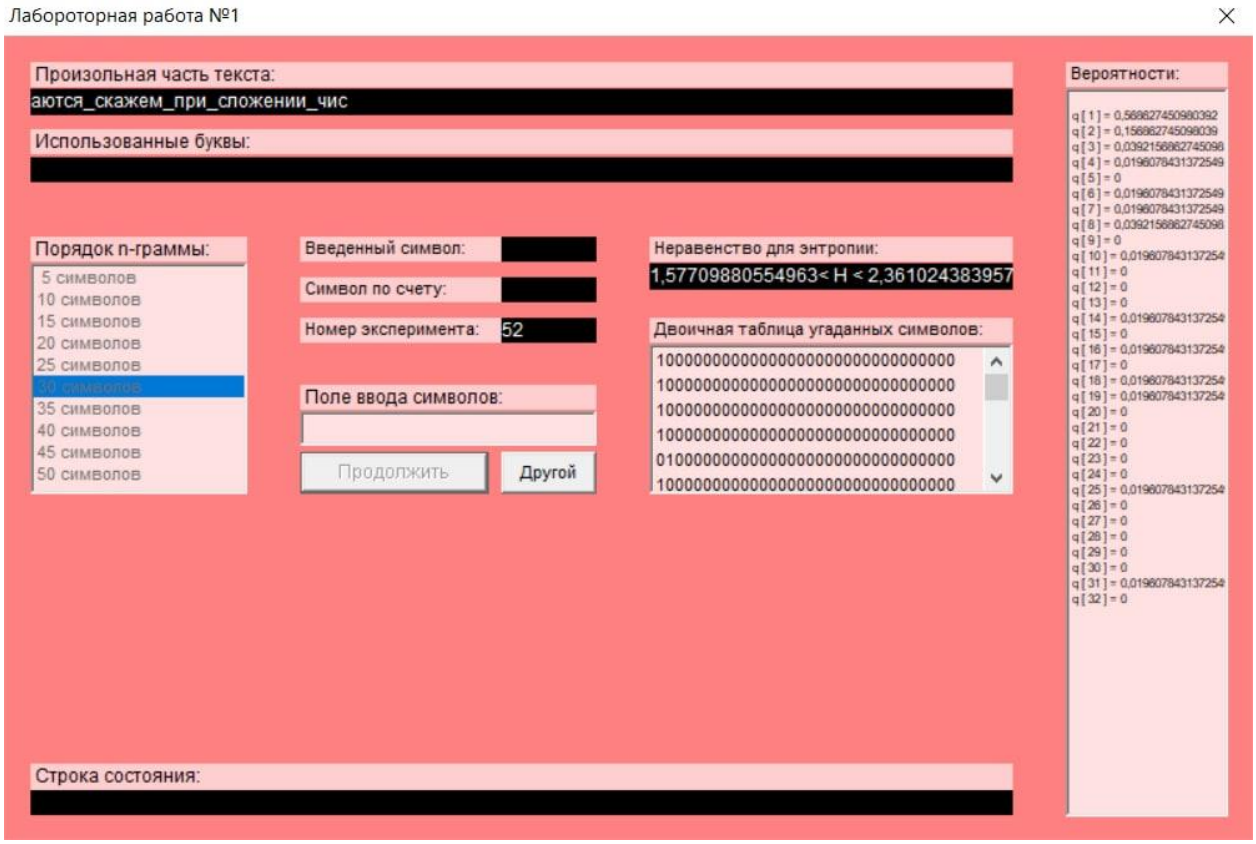
H20 = 2,583019735

R = 0,483396053



H30 = 1,96906159

R = 0,606187682



Висновок:

Найчастіший символ – пробіл, тому потрібно його прибирати при шифруванні. Найчастіші букви – “о”, “е”, “а”, “и”. Найчастіші перехресні біграми в тексті з пробілами: “о “, “е “, “и “; неперехресні: “о “, “е “, “ п“; перехресні без пробілів: “то”, “но”, “ст”; неперехресні - аналогічно.