Криптографія

Комп’ютерний практикум №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконав студент групи ФБ-11

Пташник Юрій

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 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.

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.

Виконання роботи

Зразок відкритого тексту,який використовувався в ході виконання комп'ютерного практикуму знаходиться у файлі "Prestuplenie-i-nakazanie.txt".

Спершу було написано функцію "ReadFile",яка відповідала за читання відкритого тексту з текстового файлу.Результатом виконання даної частини коду є відкритий текст у нижньому регістрі

Після цього необхідно було вирішити проблему очищення відкритого тексту від будь-яких символів,які не є літерами чи пробілами.Для цього було створено функцію "ClearText",яка повертала вже відфільтрований текст,готовий до наступної обробки.

Наступним кроком було створення функції,яка підраховує кількість різних літер в тексті.В залежності від встановленого додаткового параметра "spaces = True" функції "CountLetters" відбувався підрахунок з врахуванням пробілів.Після підрахунку і занесення даних у масив відбувається обрахунок частот,ентропії та надлишковості відкритого тексту.Це відбувається за допомогою наступних формул:

Ентропія:

$$H_1 = -\sum_{i=1}^{n} p_i \log_2 p_i$$

,де p_i – частота відповідної літери

Надлишковість:

$$R = 1 - \frac{H_1}{H_0}$$

,де H_1 – ентропія відкритого тексту,H_0 – максимальна ентропія відкритого тексту

Максимальна ентропія відкритого тексту обчислюється за формулою:

$$H_0 = \log_2 n$$

,де n – кількість букв в алфавіті

Після обчислення отримані дані записуються у відповідний csv-файл.

Далі необхідно було написати функцію,яка буде підраховувати біграми.Загалом,вона працює схожим чином з функцією для підрахунку літер,однак у “CountBigrams” присутній ще один додатковий параметр “crossing”,який відповідає за перетин біграм.Також для підрахунку ентропії використовується трохи інша формула:

$$H_2 = -\sum_{i,j} \frac{p_{i,j} \log_2 p_{i,j}}{2}$$

,де $p_{i,j}$ – частота відповідної біграми

Далі буде наведено таблиці літер та біграм з їхніми частотами.

Літери без пробілу		
Літера	Кількість літер	Частота літери
а	67149	0,079660
б	14658	0,017389
в	38996	0,046262
г	14238	0,016891
д	26989	0,032018
е	73412	0,087090
ё	65	0,000077
ж	9617	0,011409
з	12979	0,015397
и	54663	0,064848
й	8441	0,010014
к	27840	0,033027

л	38743	0,045961
м	26502	0,031440
н	54864	0,065086
о	96707	0,114725
п	23132	0,027442
р	35262	0,041832
с	44618	0,052931
т	54582	0,064752
у	24995	0,029652
ф	1049	0,001244
х	7172	0,008508
ц	2337	0,002772
ч	15260	0,018103
ш	6938	0,008231
щ	2521	0,002991
ъ	204	0,000242
ы	13920	0,016514
ь	19375	0,022985
э	2973	0,003527
ю	4735	0,005617
я	18009	0,021364

Загалом 842945 літер.

Літери з пробілом		
Літера	Кількість літер	Частота літери
	169195	0,167166
а	67149	0,066344
б	14658	0,014482
в	38996	0,038528
г	14238	0,014067
д	26989	0,026665
е	73412	0,072531
ё	65	0,000064
ж	9617	0,009502
з	12979	0,012823
и	54663	0,054007
й	8441	0,008340
к	27840	0,027506
л	38743	0,038278
м	26502	0,026184
н	54864	0,054206
о	96707	0,095547
п	23132	0,022855
р	35262	0,034839
с	44618	0,044083
т	54582	0,053927

у	24995	0,024695
ф	1049	0,001036
х	7172	0,007086
ц	2337	0,002309
ч	15260	0,015077
ш	6938	0,006855
щ	2521	0,002491
ъ	204	0,000202
ы	13920	0,013753
ь	19375	0,019143
э	2973	0,002937
ю	4735	0,004678
я	18009	0,017793

Загалом 1012140 літер.

Для таблиць біграм буде наведено лише 10 найчастіших біграм,з повними таблицями можна ознайомитися у відповідних csv-файлах або у загальній таблиці Excel.Для зручності пробіл далі буде замінятися на “_”.

Біграми без пробілів і перетинів		
Біграма	Кількість біграм	Частота біграми
то	7626	0,018093729
ов	5423	0,01286681
на	5166	0,012257042
не	5089	0,012074349
но	4992	0,011844203
ст	4897	0,011618803
по	4618	0,010956837
ко	4527	0,010740927
он	4353	0,010328088
от	4075	0,009668495

Загалом 421472 біграм.

Біграми з пробілами і без перетинів		
Біграма	Кількість біграм	Частота біграми
о_	11978	0,023668662
е_	9530	0,018831387
и_	8744	0,017278242
а_	8560	0,016914656
_в	8444	0,016685439
_н	8009	0,015825874
_п	7993	0,015794258
_с	7864	0,015539352
то	7476	0,01477266
ь_	6040	0,011935108

Загалом 506070 біграм.

Біграми без пробілів і з перетинами		
Біграма	Кількість біграм	Частота
то	15241	0,018080679
ов	10611	0,012588025
не	10313	0,012234502
на	10200	0,012100448
но	10031	0,01189996
ст	9803	0,01162948
по	9180	0,010890403
ко	8996	0,010672121
он	8747	0,010376727
от	8216	0,009746792

Загалом 842944 біграм.

Біграми з пробілами і перетинами		
Біграма	Кількість біграм	Частота
о_	23945	0,023657818
е_	19121	0,018891674
и_	17607	0,017395832
а_	17129	0,016923565
_в	16934	0,016730904
_п	16093	0,01589999
_н	16033	0,01584071
_с	15828	0,015638168
то	14871	0,014692646
ь_	12042	0,011897575

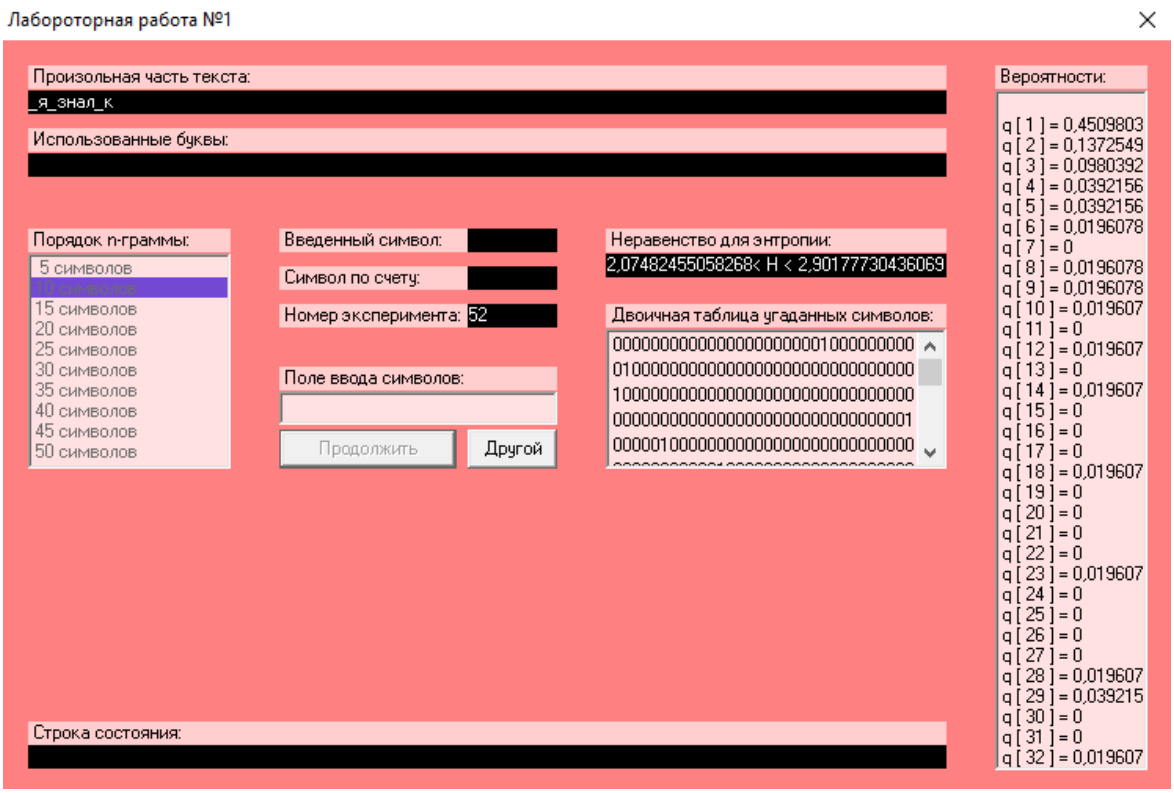
Загалом 1012139 біграм.

Далі буде наведено результати обрахування ентропії та надлишковості для кожної моделі відкритого тексту.

Модель відкритого тексту	Ентропія	Надлишковість
H_1(без пробілів)	4,45153123334034	0,117529057403134
H_1(з пробілами)	4,35856794062992	0,143272771392128
H_2(без пробілів і перетинів)	4,12649890759154	0,181963421185583
H_2(з пробілами і без перетинів)	3,94857265208392	0,223862114516498
H_2(без пробілів і з перетинами)	4,12800468949507	0,181664915187063
H_2(з пробілами і перетинами)	3,94847314637341	0,223881673521373

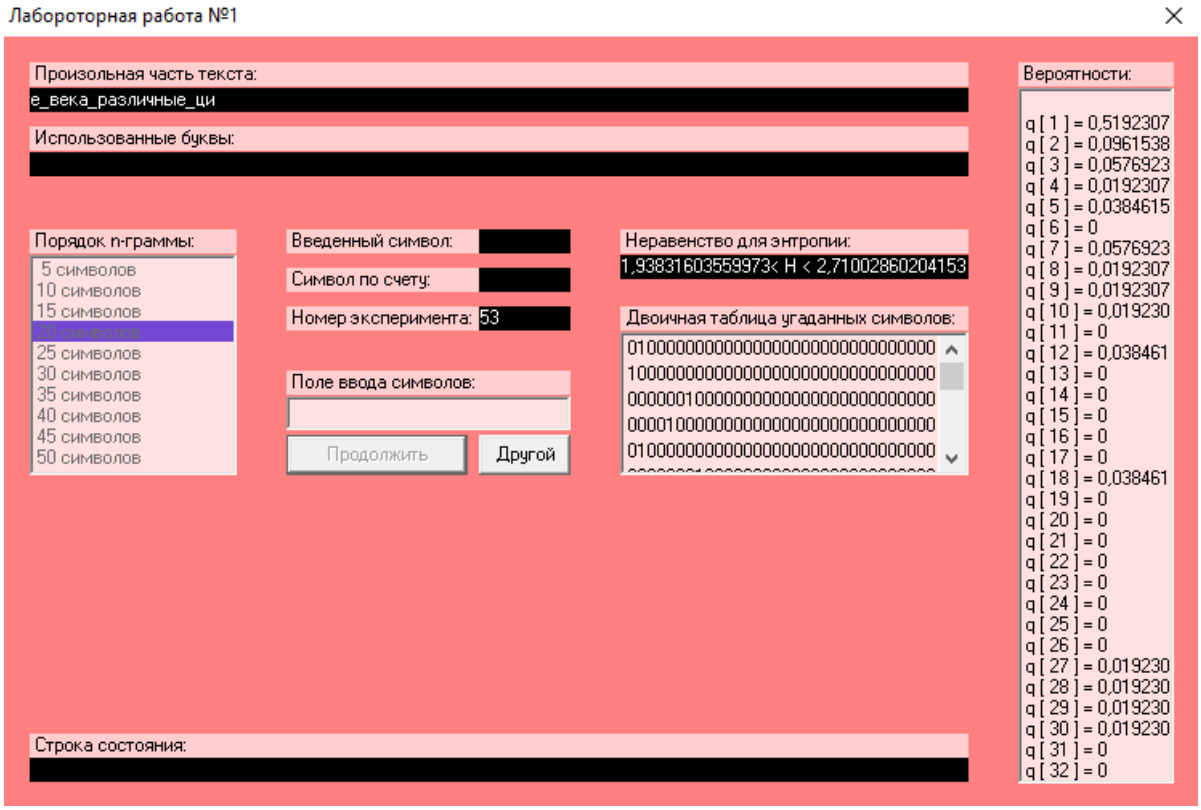
Далі за допомогою програми “CoolPinkProgram” виконаємо оцінку ентропії для n-грами,яка складається з 10,20 та 30 символів.

$$H_{10}$$



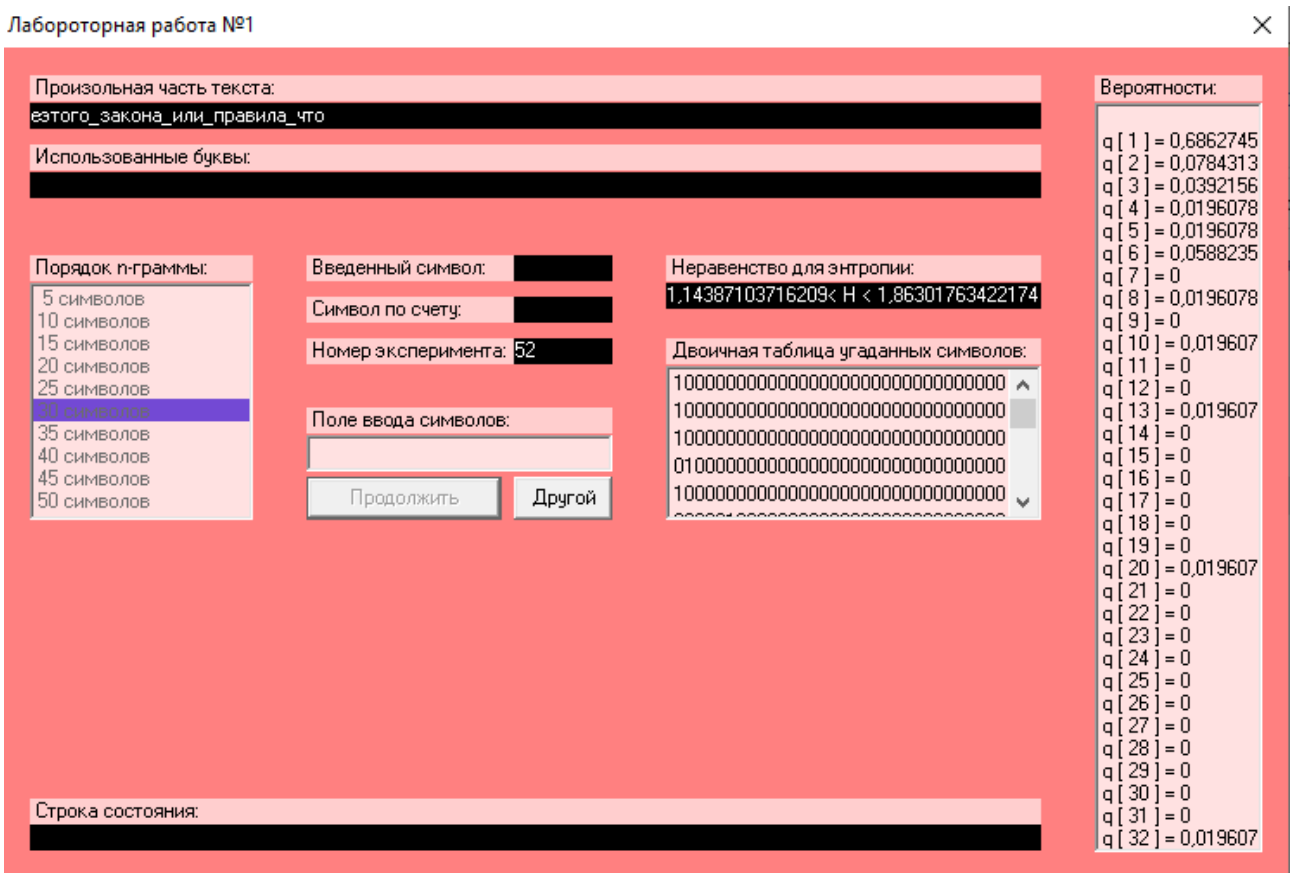
Надлишковість знаходиться в межах: $0{,}41965 < R < 0{,}58504$

H_{20}



Надлишковість знаходиться в межах: $0{,}45799 < R < 0{,}61234$

H_{30}



Надлишковість знаходиться в межах: $0{,}6274 < R < 0{,}77123$

Висновки

Виконуючи даний комп’ютерний практикум було експериментально встановлено частоту появи окремих літер та біграм для різних моделей відкритого тексту та обраховано ентропію і надлишковість таких моделей:монограми без пробілів та з пробілами;біграми без пробілів і перетинів;біграми з пробілами і перетинами;біграми лише з пробілами та біграми лише з перетинами.За допомогою програми “CoolPinkProgram” було встановлено значення ентропії та надлишковості для моделей відкритого тексту,n-грами яких складаються з 10,20 та 30 символів.