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 "sample.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 пробілів		
Літера	Кількість літер	Частота літери
а	44865	0,080017
б	7080	0,012627
в	20450	0,036473
г	10991	0,019603
д	12943	0,023084
е	44932	0,080137
ё	0	0,000000
ж	5919	0,010557
з	16699	0,029783
и	51716	0,092236
й	7908	0,014104

к	22825	0,040709
л	26243	0,046805
м	14301	0,025506
н	36811	0,065653
о	61091	0,108957
п	17280	0,030819
р	25919	0,046227
с	31093	0,055455
т	41820	0,074587
у	7156	0,012763
ф	777	0,001386
х	3910	0,006974
ц	777	0,001386
ч	5532	0,009866
ш	2420	0,004316
щ	2375	0,004236
ъ	781	0,001393
ы	5463	0,009743
ь	8724	0,015559
э	1631	0,002909
ю	7912	0,014111
я	12346	0,022019

Загалом 560690 літер.

Літери з пробілом		
Літера	Кількість літер	Частота літери
	88446	0,136252
а	44865	0,069115
б	7080	0,010907
в	20450	0,031503
г	10991	0,016932
д	12943	0,019939
е	44932	0,069218
ё	0	0,000000
ж	5919	0,009118
з	16699	0,025725
и	51716	0,079669
й	7908	0,012182
к	22825	0,035162
л	26243	0,040428
м	14301	0,022031
н	36811	0,056708
о	61091	0,094111

п	17280	0,026620
р	25919	0,039928
с	31093	0,047899
т	41820	0,064424
у	7156	0,011024
ф	777	0,001197
х	3910	0,006023
ц	777	0,001197
ч	5532	0,008522
ш	2420	0,003728
щ	2375	0,003659
ъ	781	0,001203
ы	5463	0,008416
ь	8724	0,013439
э	1631	0,002513
ю	7912	0,012189
я	12346	0,019019

Загалом 649136 літер.

Для таблиць біграм буде наведено лише 10 найчастіших біграм,з повними таблицями можна ознайомитися у відповідних csv-файлах або у загальній таблиці Excel.

Біграми без пробілів і перетинів		
Біграма	Кількість біграм	Частота біграми
ни	6515	0,023207
ст	5898	0,021009
то	5296	0,018865
пр	5179	0,018448
ос	4741	0,016888
ра	4353	0,015506
но	3738	0,013315
ко	3497	0,012456
во	3494	0,012446
ли	3346	0,011919

Загалом 280738 біграм.

Біграми з пробілами і без перетинів		
Біграма	Кількість біграм	Частота біграми
о_	8078	0,024858
_н	7108	0,021873
ни	6588	0,020273
_п	6331	0,019482
ст	5914	0,018199
то	5145	0,015833
и_	5140	0,015817

пр	5051	0,015543
ра	4354	0,013399
_в	3983	0,012257

“_” використовується замість пробілу.Загалом 324961 біграм.

Біграми без пробілів і з перетином		
Біграма	Кількість біграм	Частота біграми
ни	13075	0,023287
ст	11932	0,021251
то	10564	0,018815
пр	10264	0,018280
ос	9270	0,016510
ра	8656	0,015417
но	7531	0,013413
ко	7101	0,012647
во	7066	0,012585
ли	6735	0,011995

Загалом 561476 біграм.

Біграми з пробілом і перетином		
Біграма	Кількість біграм	Частота біграми
о_	16248	0,025000
_н	14226	0,021889
ни	13075	0,020118
_п	12471	0,019188
ст	11833	0,018207
и_	10288	0,015830
пр	10264	0,015793
то	10251	0,015773
ра	8656	0,013319
ос	7906	0,012165

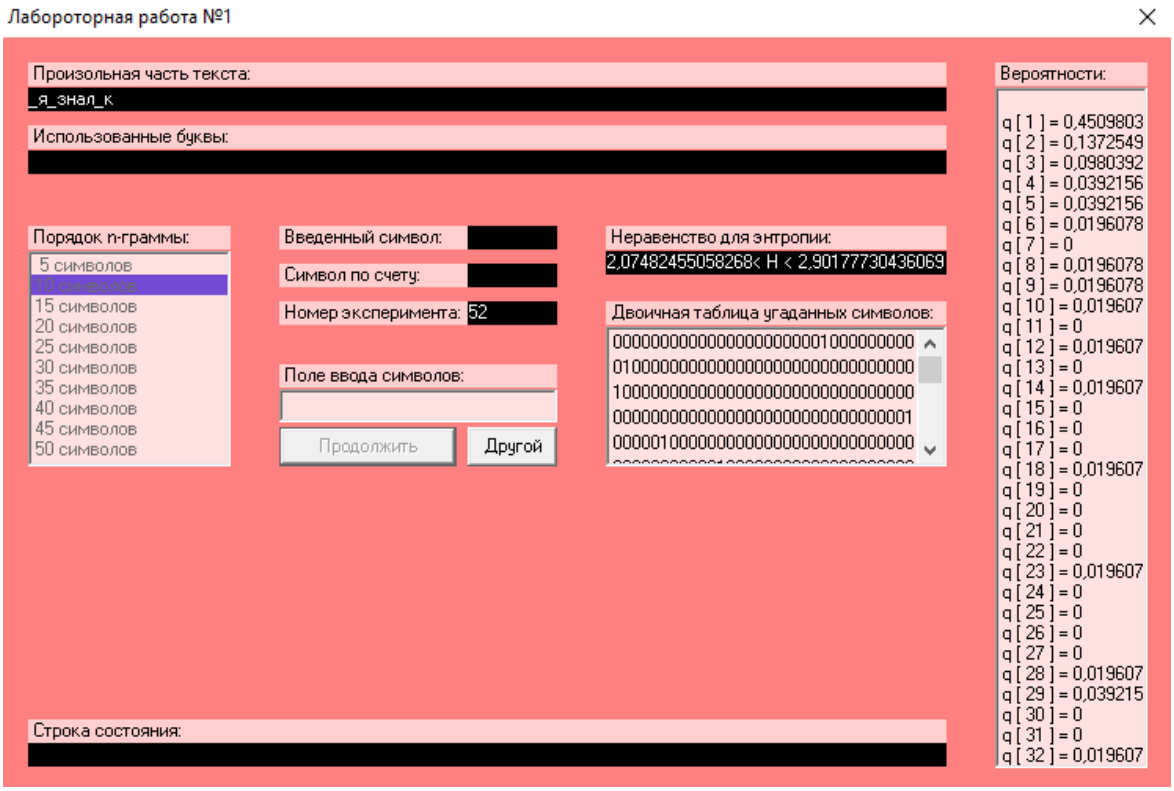
“_” використовується замість пробілу.Загалом 324961 біграм.

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

Модель відкритого тексту	Ентропія	Надлишковість
H_1(без пробілів)	4,39858	0,12803
H_1(з пробілами)	4,3736	0,14032
H_2(без пробілів і перетинів)	3,82517	0,2417
H_2(з пробілами і без перетинів)	3,70266	0,2722
H_2(без пробілів і з перетинами)	3,82496	0,24174
H_2(з пробілами і перетинами)	3,70288	0,27216

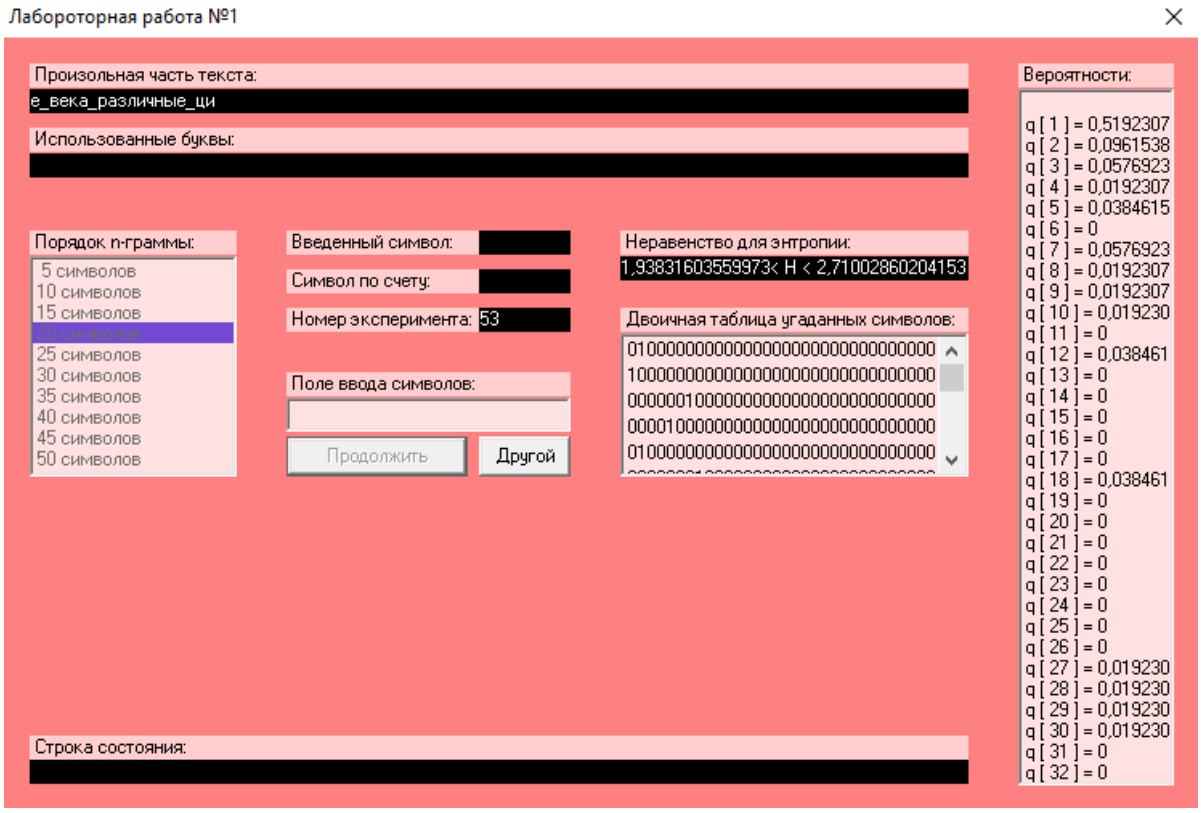
Далі за допомогою програми “CoolPinkProgram” виконаємо оцінку ентропії для n-грами,яка складається з 10,20 та 30 символів.

$$H_{10}$$



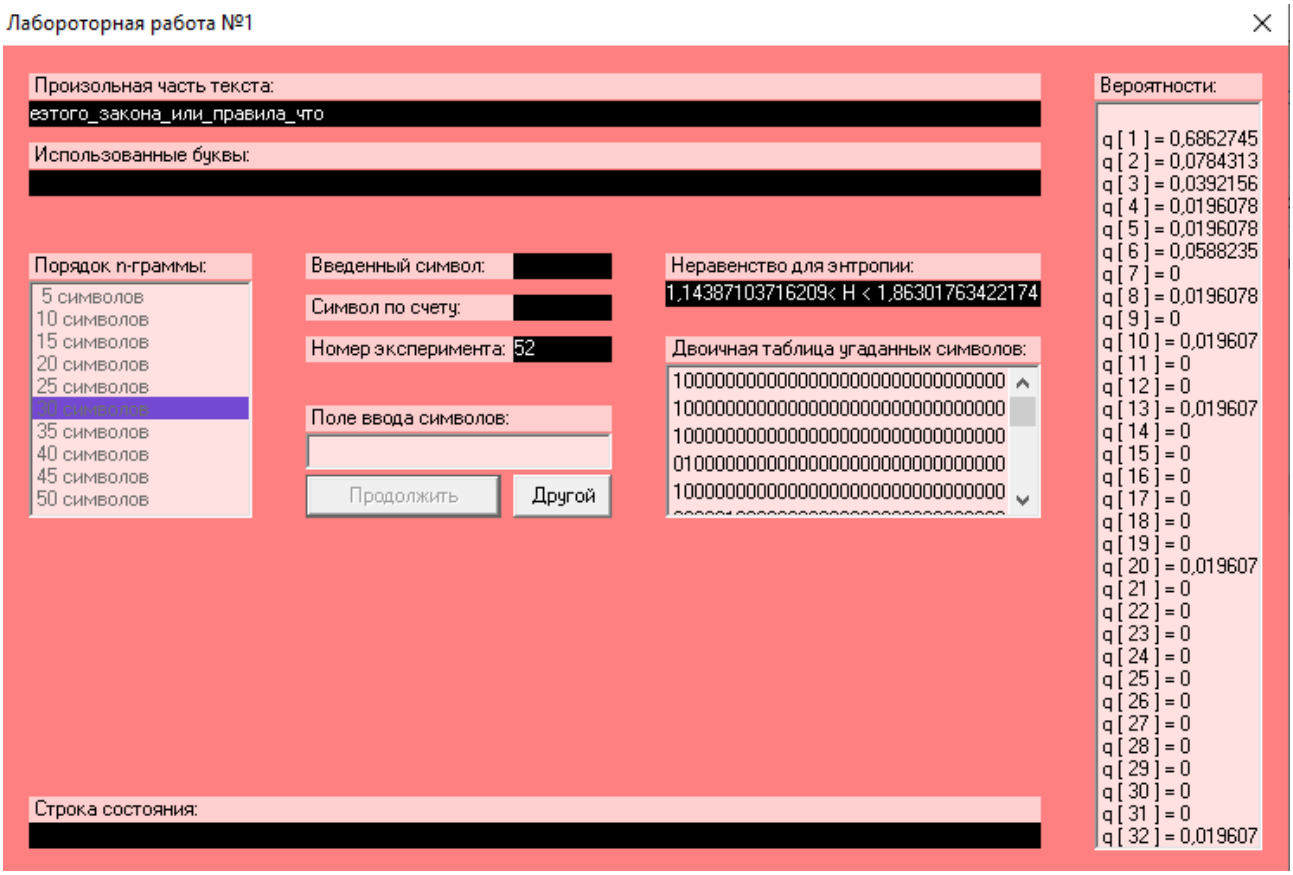
Надлишковість знаходиться в межах: $0{,}41965 < R < 0{,}58504$

$$H_{20}$$



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.