Лабораторна робота №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконали:

Ступак Ярослав ФБ-11

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

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

Хід роботи

Перед тим як почати рахувати кількість n-грам в тексті, його було необхідно очистити від зайвих символів: видалити всі знаки пунктуації, а також прибрати подвоєння пробілів якщо такі є в наявності. Написавши нескладний код я зіштовхнувся з проблемою – з'являлися додаткові пробіли в тих місцях де їх не було. Цю проблему вирішило подвійне використання коду який я використовував для очистки.

Для виконання лабораторної роботи спочатку планувалося використати сценарій фільму «Бійцівський Клуб», але нажаль його довжина виявилась замалою, тому до нього був доданий том фентезі книги (це скоріш за все трохи вплинуло на частоту появи деяких незвичних біграм)

Далі необхідно було визначити всі монограми, їх загальну кількість в тексті, та частоту появи. Я це зробив створивши словник в який додавалися літери в якості ключів і потім після кожного знаходження аналогічної літери її значення збільшувалося на 1. Після закінчення підрахунку, частота появи літери додавалася в словник. Цей самий процес аналогічний і для підрахунку біграм з однією відмінністю в тому що для визначення загальної кількості біграм був зроблений лічильник

Для обчислення H(1) застосовуємо формулу

$$H_1 = -\sum P(i) * \log_2 P(i)$$

де p(i) – частота появи літери в тексті

Для обчислення H(2) формула майже та сама

$$H_2 = \frac{-\sum P(i,j) * \log_2 P(i,j)}{2}$$

де p(i,j) – частота появи біграми

Результати обчислень

Для демонстрації біграм в таблицях бралися лише перші 30 позицій

З пробілом

Монограми		
Символ	Кількість	Частота
_	138891	0.160047
о	82048	0.094545
е	62306	0.071796
а	59862	0.06898
н	50718	0.058443
и	48141	0.055474
т	42710	0.049216
с	38985	0.044923
л	36764	0.042364
р	33507	0.038611
в	30287	0.0349
м	25006	0.028815
к	24483	0.028212
д	21717	0.025025
у	19534	0.022509
п	18574	0.021403
я	15394	0.017739
г	14454	0.016656
ь	14416	0.016612
ы	13815	0.015919
з	11854	0.01366
б	11125	0.01282
ч	10649	0.012271
й	8964	0.010329
ж	7945	0.009155
х	7220	0.00832
ш	6116	0.007048
ю	3769	0.004343
щ	2806	0.003233
ц	2505	0.002887
э	2290	0.002639
ф	852	0.000982
ъ	101	0.000116
ё	7	8.07E-06

Біграми без кроку		
Біграма	Кількість	Частота
о_	17407	0.020058
а_	15519	0.017883
е_	15243	0.017565
_с	14862	0.017126
и_	14602	0.016826

_н	14314	0.016494
_п	13725	0.015816
_в	11783	0.013578
то	10285	0.011852
на	10121	0.011663
я_	9814	0.011309
ь_	9468	0.01091
_о	9436	0.010873
ст	9236	0.010643
не	8687	0.01001
но	8526	0.009825
_к	8166	0.00941
ко	8050	0.009276
по	8012	0.009232
_и	7972	0.009186
ал	7241	0.008344
_м	6850	0.007893
ро	6751	0.007779
_т	6709	0.007731
ни	6574	0.007575
ол	6350	0.007317
й_	6298	0.007257
го	6274	0.00723
ен	6264	0.007218
ра	6139	0.007074

Біграми з кроком		
Біграма	Кількість	Частота
о_	17407	0.020058
а_	15519	0.017883
е_	15243	0.017565
_с	14862	0.017126
и_	14602	0.016826
_н	14314	0.016494
_п	13725	0.015816
_в	11783	0.013578
то	10285	0.011852
на	10121	0.011663
я_	9814	0.011309
ь_	9468	0.01091
_о	9436	0.010873
ст	9236	0.010643
не	8687	0.01001
но	8526	0.009825
_к	8166	0.00941
ко	8050	0.009276
по	8012	0.009232
_и	7972	0.009186
ал	7241	0.008344
_м	6850	0.007893
ро	6751	0.007779
_т	6709	0.007731
ни	6574	0.007575
ол	6350	0.007317

й_	6298	0.007257
го	6274	0.00723
ен	6264	0.007218
ра	6139	0.007074

Без пробілів

Монограми		
Символ	Кількість	Частота
о	82048	0.11256
е	62306	0.085477
а	59862	0.082124
н	50718	0.069579
и	48141	0.066044
т	42710	0.058593
с	38985	0.053483
л	36764	0.050436
р	33507	0.045968
в	30287	0.04155
м	25006	0.034305
к	24483	0.033588
д	21717	0.029793
у	19534	0.026798
п	18574	0.025481
я	15394	0.021119
г	14454	0.019829
ь	14416	0.019777
ы	13815	0.018953
з	11854	0.016262
б	11125	0.015262
ч	10649	0.014609
й	8964	0.012298
ж	7945	0.0109
х	7220	0.009905
ш	6116	0.00839
ю	3769	0.005171
щ	2806	0.00385
ц	2505	0.003437
э	2290	0.003142
ф	852	0.001169
ъ	101	0.000139
ё	7	9.6E-06

Біграми без кроку		
Біграма	Кількість	Частота
то	10582	0.014517
на	10151	0.013926
ст	9485	0.013012
но	8765	0.012025
не	8743	0.011994
ко	8272	0.011348
по	8017	0.010998

он	7669	0.010521
ен	7640	0.010481
ал	7528	0.010328
ос	7070	0.009699
ов	7006	0.009611
ни	6853	0.009402
ро	6824	0.009362
ол	6710	0.009205
ор	6425	0.008814
го	6331	0.008685
от	6226	0.008541
ра	6157	0.008447
ер	6110	0.008382
ли	6054	0.008305
ес	5724	0.007853
ла	5538	0.007598
ан	5419	0.007434
ка	5275	0.007237
ом	5188	0.007117
пр	5156	0.007073
ло	5131	0.007039
ас	5077	0.006965
та	5053	0.006932

Біграми з кроком		
Біграма	Кількість	Частота
то	5339	0.014649
на	5063	0.013892
ст	4723	0.012959
не	4389	0.012042
но	4376	0.012007
ко	4192	0.011502
по	4020	0.01103
ен	3865	0.010605
он	3821	0.010484
ал	3654	0.010026
ос	3576	0.009812
ов	3476	0.009537
ни	3435	0.009425
ро	3397	0.009321
ол	3361	0.009222
ор	3213	0.008816
го	3163	0.008679
от	3098	0.0085
ра	3083	0.008459
ер	3080	0.008451
ли	3072	0.008429
ес	2863	0.007855
ла	2783	0.007636
ка	2682	0.007359
ан	2663	0.007307
ом	2600	0.007134
та	2572	0.007057
пр	2569	0.007049

ло	2532	0.006947
ть	2498	0.006854

Далі рахуємо надлишковість за формулою

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

Для тексту з пробілами $H_0 = 5.087$

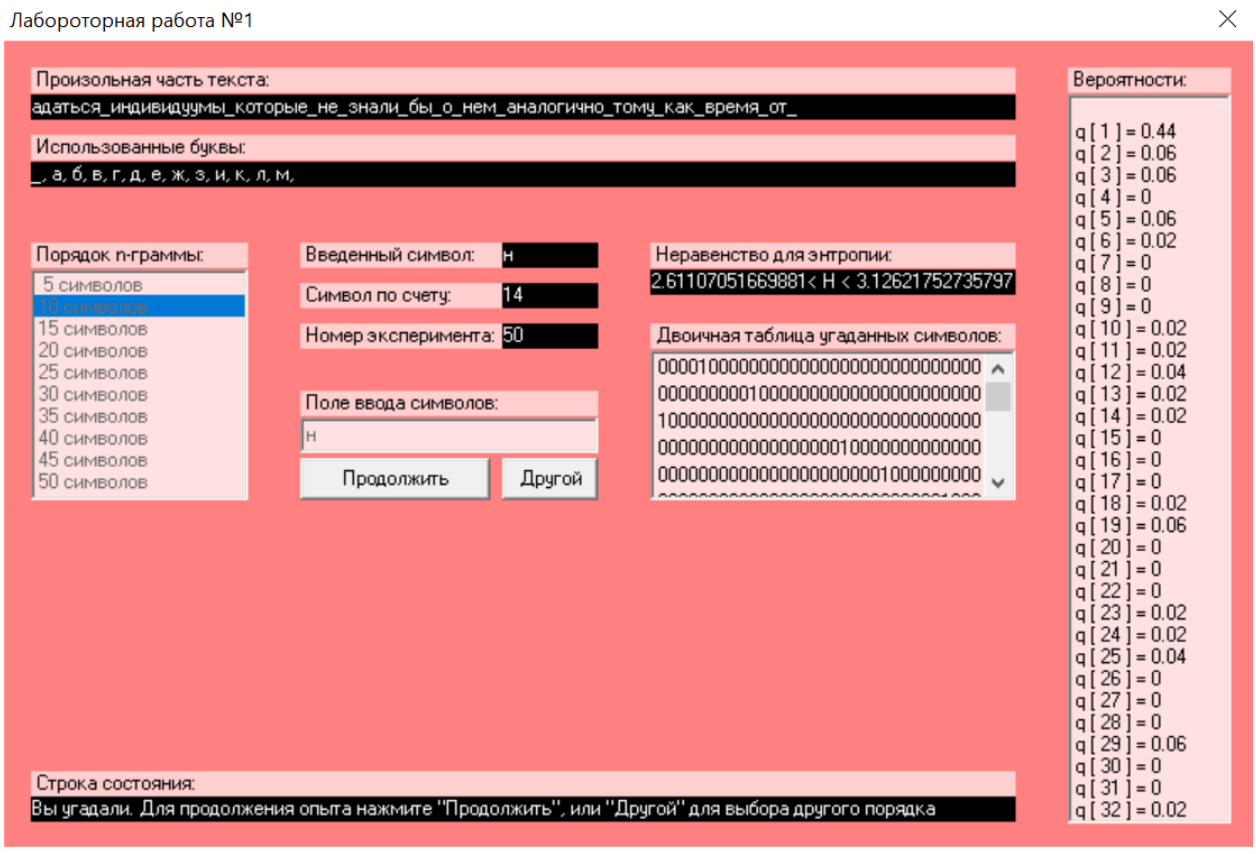
Для тексту без пробілів $H_0 = 5.044$

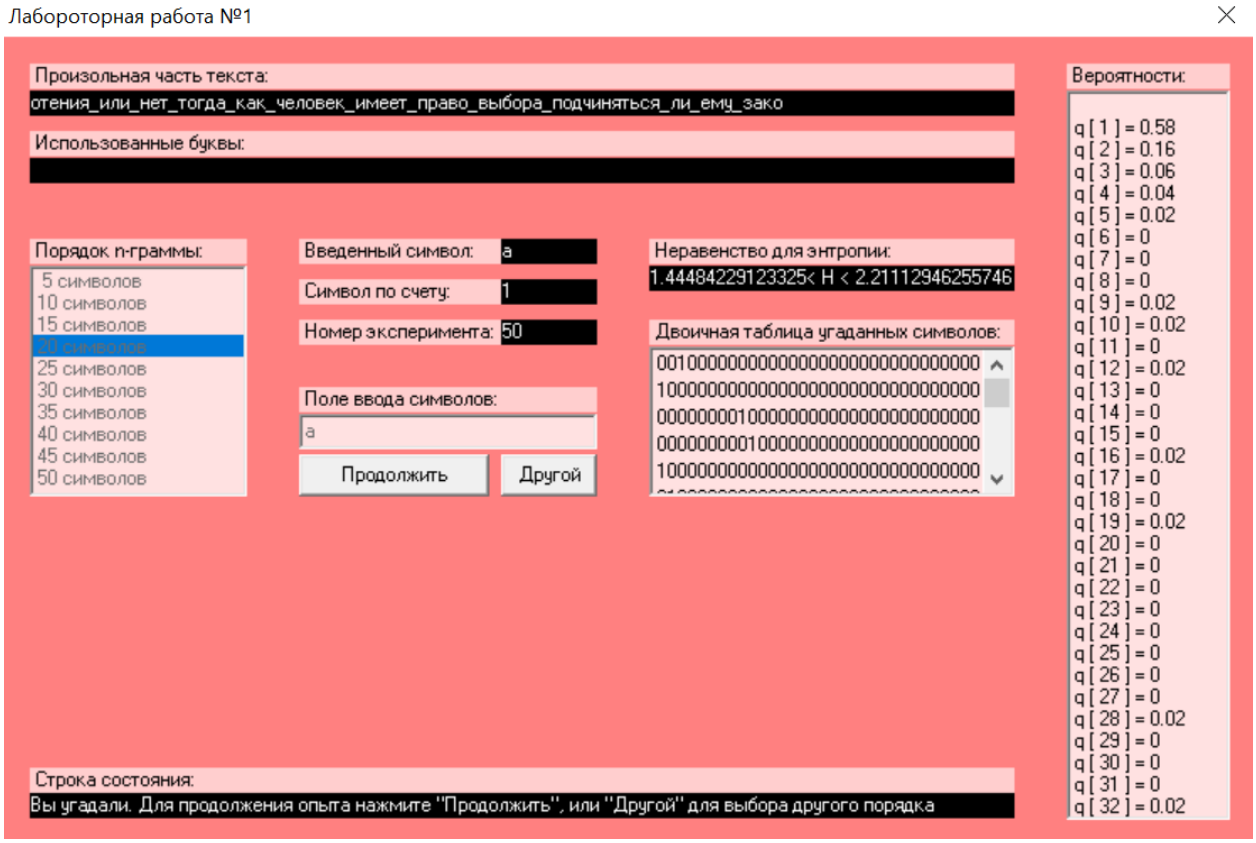
	Ентропія	Надлишковість
Монограми	4.375885080541638	0.139791
Монограми без пробілів	4.45437143943164	0.116897
Біграми	3.979607056234021	0.217691
Біграми без пробілів	4.1454607359049715	0.17814
Біграми з кроком	3.979382315524375	0.217735
Біграми з кроком без пробілів	4.1449427793149765	0.178243

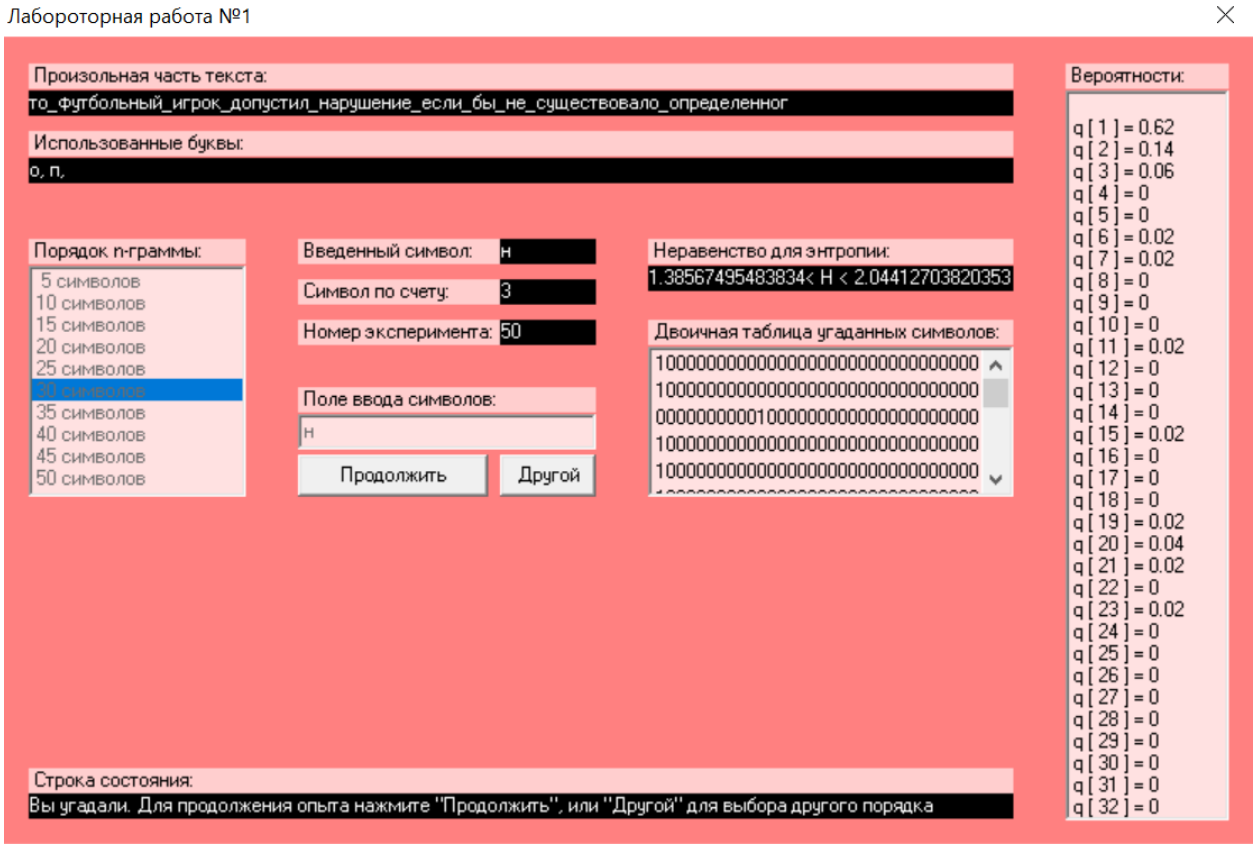
$H_{10} = 2.61 < H < 3.12$

$H_{20} = 1.44 < H < 2.21$

$H_{30} = 1.38 < H < 2.04$







Висновки

Виконавши цю лабораторну роботу ми самостійно визначили які літери та біграми зустрічаються в російській мові частіше за інших, що допоможе нам при аналізи

шифрованого тексту. Також ми змогли напряму побачити залежність між довжиною n-грами та ентропією.