КРИПТОГРАФІЯ КОМП'ЮТЕРНИЙ ПРАКТИКУМ №1 Експериментальна оцінка ентропії на символ джерела відкритого тексту

ФБ-12 Карабінський Василь, Мосейко Олег

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.

Для початку треба відфільтрувати текстовий файл, для цього створимо дві функції `filtrate(ini_text, write_file)` та `filtrate_with_spaces(ini_text, write_file)` ці жві функції очищують наш файд від зайвих символів та переписують очищений зміст у новий файл, різниця між ними тільки в тому, що перша функція прибирає пробіли, а друга ні, отже на виході ми отримали два текстові файли lab1CP_filtered.txt та lab1CP_filtered_spaces.txt

Далі за допомогою функцій `calculate_letter_frequencies(write_file)` та `total_letter_freq(letter_frequencies, file)` порахуємо частоти букв у тексті для тексту з пробілами та без пробілів
отримали такі значення
Частоти букв без пробілів:

буква	кількість	частота
а	67149	0.07966
б	14658	0.01739
в	38996	0.04626
г	14238	0.01689
д	26989	0.03201
е	73412	0.08709
ё	0	0
ж	9617	0.01140
з	12979	0.01539
и	54663	0.06485
й	8441	0.01001
к	27840	0.03302
л	38743	0.04596

м	26502	0.03144
н	54864	0.06509
о	96707	0.11473
п	23132	0.02744
р	35262	0.04183
с	44618	0.05293
т	54582	0.06475
у	24995	0.02965
ф	1049	0.00124
х	7172	0.00850
ц	2337	0.00277
ч	15260	0.01810
ш	6938	0.00823
щ	2521	0.00299
ъ	204	0.00024
ы	13920	0.01651
ь	19375	0.02298
э	2973	0.00352
ю	4735	0.00561
я	18009	0.02136

Частоти літер з пробілами

буква	кількість	частота
а	67149	0.06605
б	14658	0.01441
в	38996	0.03835
г	14238	0.01400
д	26989	0.02654
е	73412	0.07221
ё	0	0
ж	9617	0.00945
з	12979	0.01276
и	54663	0.05377
й	8441	0.00830
к	27840	0.02738

л	38743	0.03811
м	26502	0.02606
н	54864	0.02606
о	96707	0.09512
п	23132	0.02275
р	35262	0.03468
с	44618	0.04388
т	54582	0.05369
у	24995	0.02458
ф	1049	0.00103
х	7172	0.00705
ц	2337	0.00229
ч	15260	0.01501
ш	6938	0.00682
щ	2521	0.00247
ъ	204	0.00020
ы	13920	0.01369
ь	19375	0.01905
э	2973	0.00292
ю	4735	0.00465
я	18009	0.01771

_	173725	0.17088

Далі за допомогою функції `calculate_bigram_frequencies(writefile)` знайдемо частоти біграм для початку біграми без пробілів з повторами:

біграма	кількість	частота
то	15241	0.01808
ов	10611	0.01258
не	10313	0.01223
на	10200	0.01210
но	10031	0.01190
ст	9810	0.01163
по	9180	0.01089
ко	8996	0.01067
он	8747	0.01037

от	8216	0.00974
ен	8106	0.00961
ни	7890	0.00936
ос	7878	0.00934
го	7779	0.00922
ал	7445	0.00883
ра	7400	0.00877
ро	7249	0.00860
ка	6986	0.00828
пр	6751	0.00800

біграми з пробілами з повторами

біграма.	кількість	частота
о_	24826	0.02442
е_	19701	0.01937
и_	17870	0.01757
а_	17526	0.01723
_в	17078	0.01679
_н	16553	0.01628
_с	16532	0.01628
_п	16278	0.01601
то	14862	0.01461
ь_	12318	0.01211
_и	11832	0.01163
я_	11761	0.01156
_о	11678	0.01148
_т	10531	0.01148
не	10230	0.01006
на	10149	0.00998
но	9757	0.00959
ст	9516	0.00936
по	9178	0.0090

біграми без пробілів без повторів

біграми.	кількість	частота
то	7532	0.01787
ов	5348	0.01268
на	5136	0.01218
не	5064	0.01201
но	4962	0.01177
ст	4860	0.01153
по	4518	0.01072
ко	4462	0.01058
он	4429	0.01050
от	4170	0.00989
ен	4083	0.00968
го	3957	0.00938
ни	3898	0.00924
ос	3880	0.00920
ра	3722	0.00883
ро	3697	0.00877
ал	3652	0.00866
ка	3475	0.00824
пр	3394	0.00805

біграми без повторів з пробілами

біграми	кількість	частота
о_	12358	0.02427
е_	9918	0.01948
и_	8945	0.01757
а_	8611	0.01691
_в	8573	0.01684
_н	8315	0.01633
_с	8188	0.01608
_п	8093	0.01589

то	7430	0.01459
ь_	6160	0.01210
_и	5932	0.01165
_о	5881	0.01155
я_	5861	0.01151
на	5201	0.01021
_т	5144	0.01010
не	5049	0.00991
ст	4902	0.00963
но	4864	0.00955
по	4639	0.00911

Тепер можемо порахувати ентропію для кожного з випадків, рахуємо за допомогою функції `calculate_entropy` та формули:

$$H(x_1, x_2, ..., x_n) = -\sum_{z_1, z_2, ..., z_n} P(x_1 = z_1, ..., x_n = z_n) \cdot \log_2 P(x_1 = z_1, ..., x_n = z_n).$$

1. ентропія для тексту без пробілів з повторами
   H1 = 4.450709647518523
   H2 = 4.127358996007625

2. з повторами з пробілом
   H1 = 3.9142983246663468. виправлено: H1 = 4.349870170166867
   H2 = 3.936494657306602

3. H2 без повторів без пробілів \/ H2 без повторів, з пробілами
   ентропія H2 для тексту без пробілів без повторів = 4.130939454893129
   ентропія H2 для тексту з пробілами без повторів = 3.938820765809703

як ми бачимо у пункті 2 H1 пораховано неправильно, провівши аналіз ми знайшли в чому заключалася помилка. наша функція для підрахунку частоти символів, не врахувала пробіли, через це значення H1 вийшло менше ніж повинно було б, тому ми додали в словник, який містить значення частоти запис з частотою пробілів та порахували H1 заново, ось, що з цього вийшло:

```
file_to_count = open(write_file_spaces, 'r', encoding='utf-8')
data_to_count = file_to_count.read()
tlf_spaces[' '] = letter_frequencies_space[' ']/ len(data_to_count)
```

```
ентропія H1 для тексту без пробілів з повторами  4.450709647518523
ентропія H2 для тексту без пробілів з повторами  4.127358996007625
ентропія H1 для тексту з пробілами з повторами  4.349870170166867
ентропія H2 для тексту з пробілами з повторами  3.936494657306602
ентропія H2 для тексту без пробілів без повторів  4.130939454893129
ентропія H2 для тексту з пробілами без повторів 3.938820765809703
```

Як можемо бачити помилку випралено і програма рахує ентропію H1 правильно.

2.Порахуємо надлишковість за формулою:

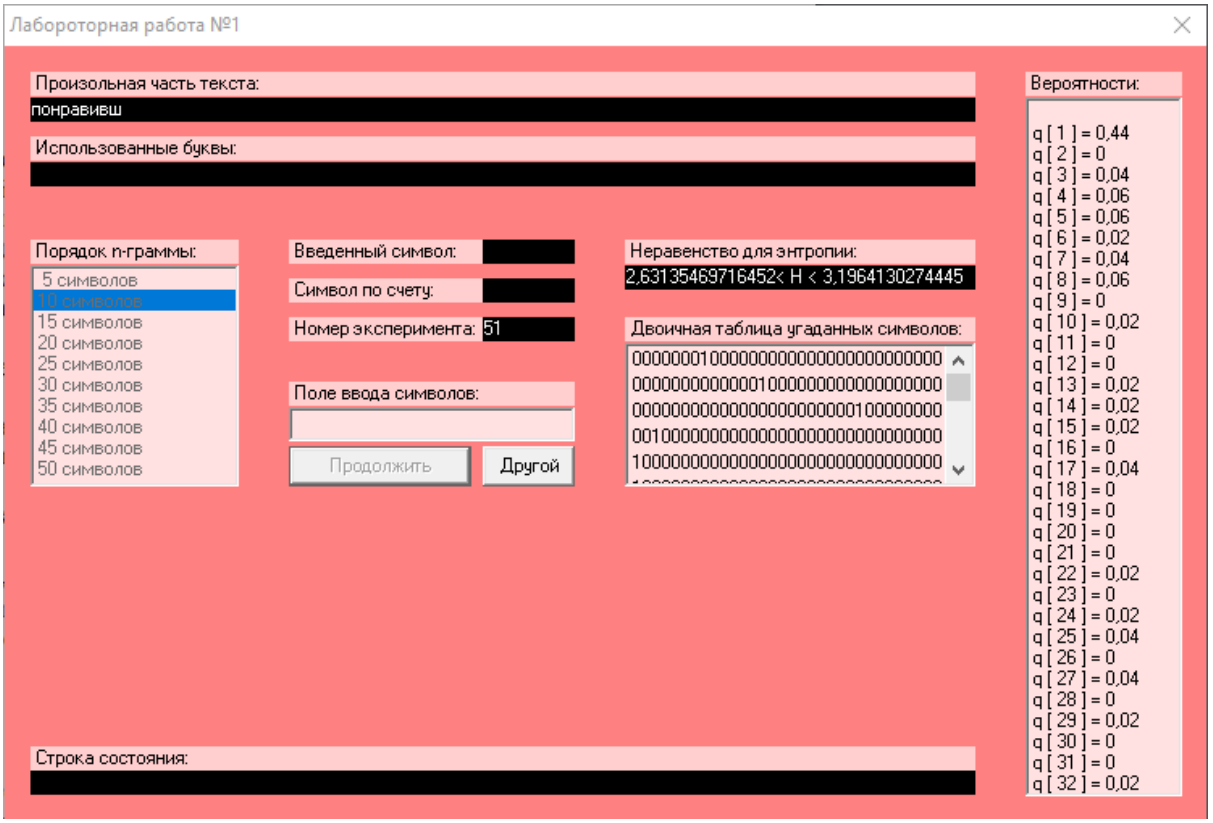
$$R = 1 - \frac{H_\infty}{H_0}$$

за H(8) в нас будуть ті H1 nf H2 що ми знайшли, а H0 = log2(33) - для тексту без пробілів та log2(34) для тексту з пробілами

```
надлишковість для H1 без пробілів 0.11769192846403431
надлишковість для H1  з пробілами 0.23059913225737982
надлишковість для H2 без пробілів з повторами 0.18179291737566639
надлишковість для H2 з пробілами з повторами 0.22623618488402852
надлишковість для H2 без пробілів без повторів 0.18108312769611612
надлишковість для H2 з пробілами без повторів 0.22577896119990837
```

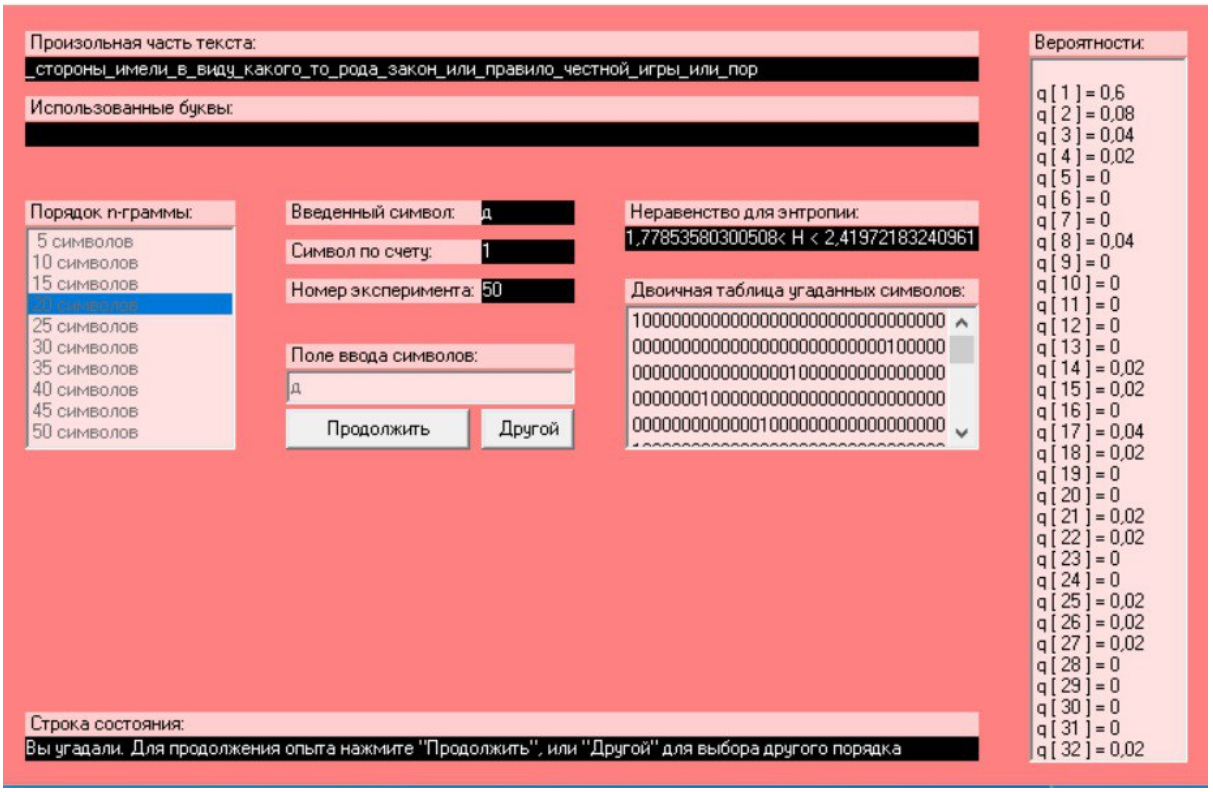
3. За допомогою програми CoolPinkProgram оцінити значення (10) H , (20) H , (30) H

3.1 H(10)



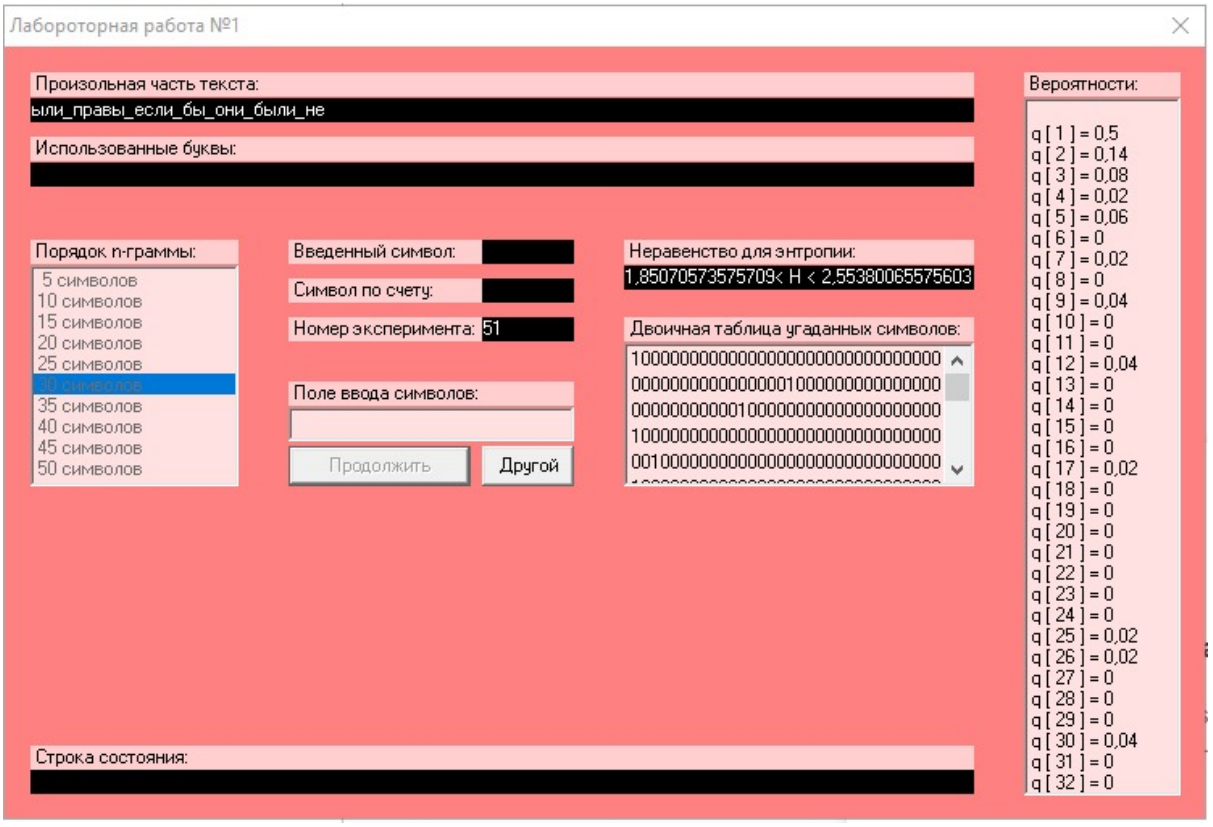
0.48277662574183733 < R10 < 0.371707838035251

3.2 H(20)



0.6504080995768076 < R20 < 0.5243755270721706

3.3 H(30)



0.6362222597969395 < R30 < 0.49802077470734196

Висновок

В ході виконнання даної лабораторної роботи ми засвоїли поняття ентропії на символ джерела та надлишковості, набули практичних навичок для оцінки ентропії, також скориштувавшись coolpinkprogram та таблицею з біграмами з найбільшими частотами навчились вгадувати наступний символ в реченні, навіть якщо цей символ стоїть на початку.