КРИПТОГРАФІЯ

КОМП'ЮТЕРНИЙ ПРАКТИКУМ №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконали:

Винник Михайло та Кузнєцов Олексій ФБ-12

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 H1() 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:

1. Спочатку в коді програми ми відфільтрували текст та очистили його від зайвих символів, знаків переносу слова, подвійних пробілів, пробілів взагалі та чисел та інших символів, що не належать російському алфавіту. Таким чином отримуємо другу версію нашого текстового документу. Відповідний крок продемонстровано в коді методом *filter()*.
2. Далі починається виконання основного завдання:

В наступній частині програми ми виконуємо задачу, яка відповідає підрахунку частот букв, а потім і частот біграм в тексті. Щоб це здійснити, ми спочатку рахуємо кількість, яку трапляється кожна літера з алфавіту в нашому тексті. Також рахуємо кількість пробілів, тому що будемо порівнювати два випадки: коли в нашому алфавіті є пробіл, та коли його нема. Для цієї задачі використовуємо словники. Вони дуже зручні для того, щоб зберігати інформацію про елемент, та його кількість.

Після того, як ми підрахували, скільки в тексті трапляється кожна літера, ми можемо порахувати імовірність, або ж частоту: ділимо кількість кожного окремого символу алфавіту на загальну кількість символів в тексті, що відповідають символам алфавіту.

3. Тепер можна легко обчислити H(1). Для цього скористаємось формулою:

$$H_1 = -\sum_{i=1}^{n} p(i) \log_2 p(i)$$

Де n – кількість літер алфавіту, p(i) – імовірність (частота) появи літери в тексті.

Цю формулу ми теж реалізували в своєму коді.

Далі обчислюємо H(2). Для цього треба порахувати частоту біграм. Робимо це так само, як і з частотою символів, але з урахуванням того, що частота біграм – відношення кількості поява деякої біграми до загальної кількості біграм у тексті. Щоб порахувати біграми, ми розробили фрагмент коду, що їх формує з тексту та рахує в словнику.

4. H(2) обчислюємо за такою формулою:

$$H_2 = -\sum_{i,j} p(i,j) \log_2 p(i,j)/2$$

де p(i,j) – частота появи деякої біграми в тексті.

Таку формулу ми отримуємо з загальної формули для розрахунку ентропії n-грам:

$$H_n = \frac{1}{n} * H(x_1, x_2, x_3 \dots, x_n)$$

Де $H(x_1, x_2, x_3 \dots, x_n)$ в свою чергу, – ентропія n-грами відкритого тексту $(x_1, x_2, x_3 \dots, x_n)$.

Результати виконання програми:

(на наступній сторінці)

Таблиця частот літер

З пробілом:

Буква	Кількість	Частота
а	47917	0.06147
б	10629	0.01363
в	29047	0.03726
г	11416	0.01464
д	19428	0.02492
е	56840	0.07291
ё	1	0.0
ж	6235	0.008
з	12203	0.01565
и	49448	0.06343
й	7911	0.01015
к	20658	0.0265
л	31596	0.04053
м	23876	0.03063
н	45985	0.05899
о	73763	0.09462
п	18248	0.02341
р	29569	0.03793
с	36491	0.04681
т	37317	0.04787
у	17100	0.02194
ф	1448	0.00186
х	7790	0.00999
ц	1981	0.00254
ч	9675	0.01241
ш	5103	0.00655
щ	3072	0.00394
ъ	150	0.00019
ы	14333	0.01839
ь	11523	0.01478
э	2296	0.00295
ю	3876	0.00497
я	14067	0.01804
_	118574	0.1521

Частота "ё": 1.2827650256681283e-06

Таблиця частот літер

Без пробіла:

Буква	Кількість	Частота
а	47917	0.07249
б	10629	0.01608
в	29047	0.04394
г	11416	0.01727
д	19428	0.02939
е	56840	0.08599
ё	1	0.0
ж	6235	0.00943
з	12203	0.01846
и	49448	0.07481
й	7911	0.01197
к	20658	0.03125
л	31596	0.0478
м	23876	0.03612
н	45985	0.06957
о	73763	0.11159
п	18248	0.02761
р	29569	0.04473
с	36491	0.05521
т	37317	0.05646
у	17100	0.02587
ф	1448	0.00219
х	7790	0.01179
ц	1981	0.003
ч	9675	0.01464
ш	5103	0.00772
щ	3072	0.00465
ъ	150	0.00023
ы	14333	0.02168
ь	11523	0.01743
э	2296	0.00347
ю	3876	0.00586
я	14067	0.02128

Частота "ё": 1.5128776142525175e-06

Таблиця частот біграм з повторами

З пробілом:

Біграма	Кількість	Частота
и_	15616	0.02003
о_	14985	0.01922
е_	13857	0.01778
_п	12465	0.01599
_с	11913	0.01528
_в	11166	0.01432
_н	11063	0.01419
а_	9498	0.01218
ст	9300	0.01193
_и	8901	0.01142
я_	8746	0.01122
но	8703	0.01116
то	8703	0.01116
_о	8700	0.01116
не	7405	0.0095
на	7334	0.00941
по	7285	0.00934
ни	6967	0.00894
ь_	6875	0.00882
м_	6600	0.00847

Таблиця частот біграм з піторами

Без пробілу:

Біграма	Кількість	Частота
ст	9436	0.01428
то	8912	0.01348
но	8871	0.01342
ен	7677	0.01161
не	7429	0.01124
на	7349	0.01112
по	7287	0.01102
ни	7127	0.01078
ос	6824	0.01032
ов	6672	0.01009
ли	6631	0.01003
ра	6469	0.00979
ко	6323	0.00957
ал	6183	0.00935
ро	5582	0.00844
ес	5487	0.0083
го	5480	0.00829
ом	5474	0.00828
ер	5451	0.00825
он	5316	0.00804

Таблиця частот біграм без повторів

З пробілом:

Біграма	Кількість	Частота
и_	7736	0.01985
о_	7522	0.0193
е_	6956	0.01785
_п	6283	0.01612
_с	5992	0.01537
_н	5575	0.0143
_в	5573	0.0143
а_	4767	0.01223
ст	4548	0.01167
_и	4478	0.01149
я_	4444	0.0114
то	4380	0.01124
_о	4344	0.01114
но	4279	0.01098
не	3704	0.0095
на	3683	0.00945
по	3592	0.00922
ь_	3404	0.00873
ни	3399	0.00872
м_	3313	0.0085
ен	3281	0.00842
ра	3202	0.00821

Таблиця частот біграм без повторів

Без пробіла:

Біграма	Кількість	Частота
ст	4670	0.01413
но	4531	0.01371
то	4443	0.01344
ен	3788	0.01146
не	3742	0.01132
по	3663	0.01108
ни	3654	0.01106
на	3617	0.01094
ос	3389	0.01025
ов	3289	0.00995
ли	3266	0.00988
ра	3151	0.00953
ко	3134	0.00948
ал	3097	0.00937
ро	2845	0.00861
ес	2794	0.00845
ом	2790	0.00844
ер	2761	0.00835
го	2706	0.00819
от	2674	0.00809
он	2662	0.00805
ре	2583	0.00782

Надлишковість відкритого тексту обчислюється за формулою:

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

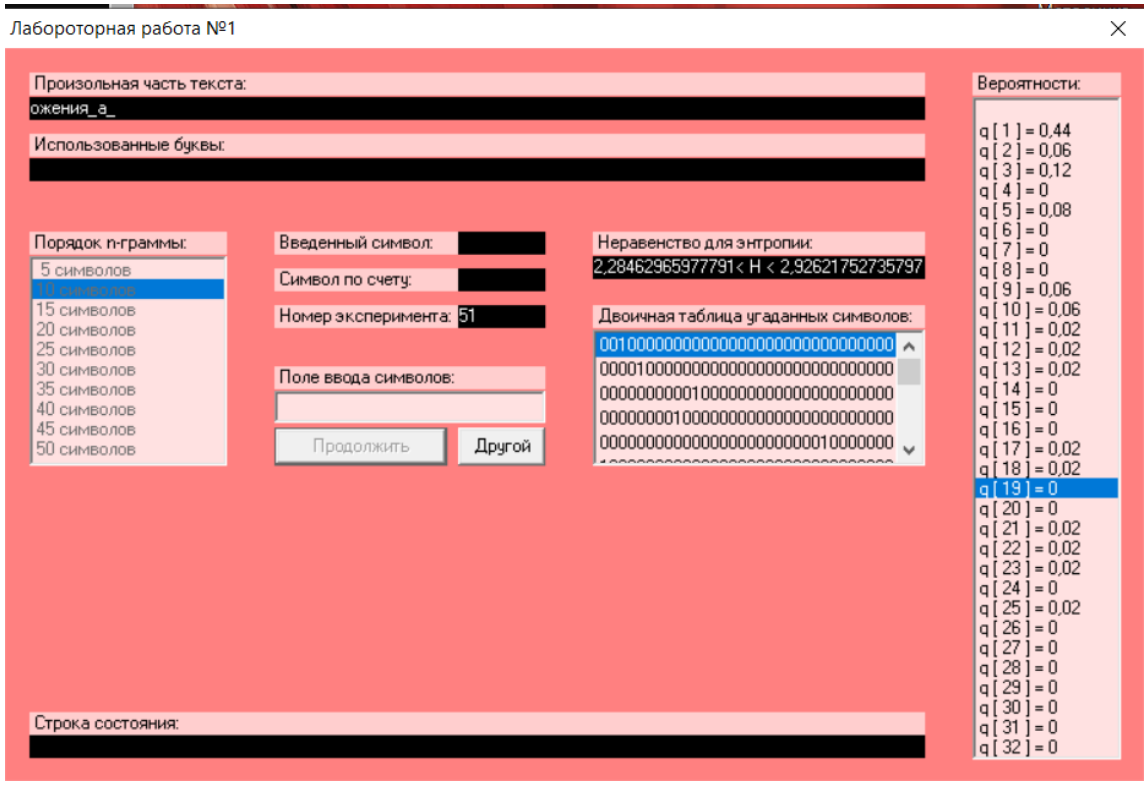
$H_0 = \log_2 34 = 5.08746$ для тексту з пробілами, $\log_2 33 = 5.04439$ для тексту без пробілів

	Ентропія	Надлишковість
H1 (з пробілом)	4,4032187	0,134417397
H1 (без пробіла)	4,4676885	0,11425684
H2 (з пробілом, з повторами)	4,0015652	0,213374248
H2 (без пробіла, з повторами)	4,1504491	0,177151249
H2 (з пробілом, без повторів)	4,0017929	0,213329487
H2 (без пробіла, без повторів)	4,1496428	0,177311102

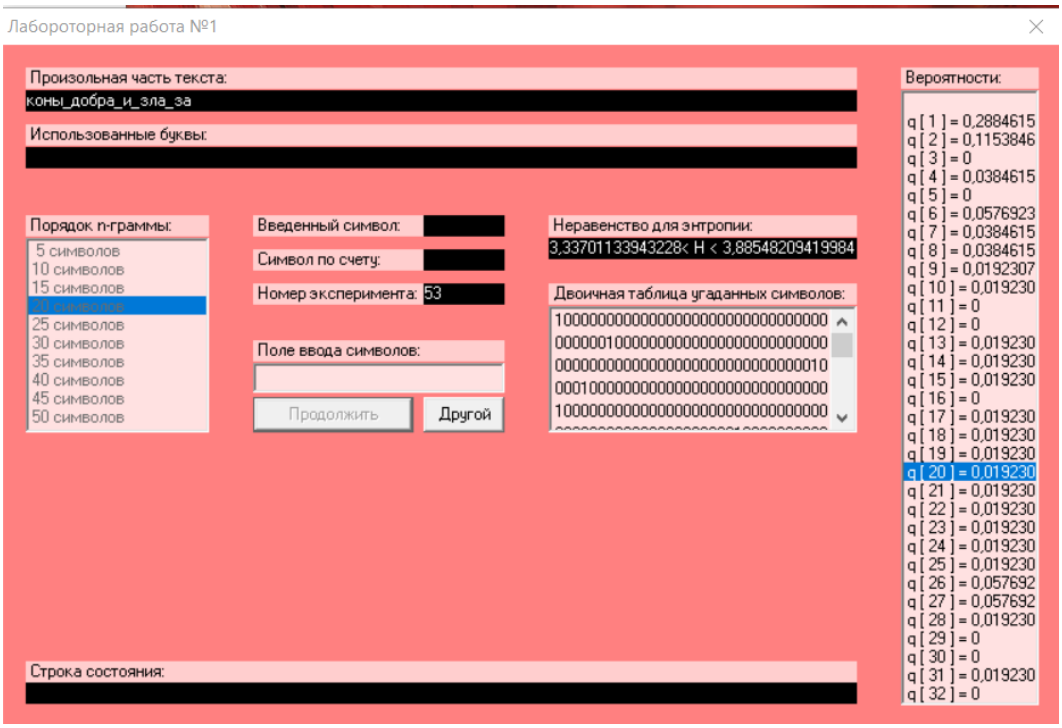
Експеременти в програмі CoolPinkProgram

	Ентропія	Надлишковість
H(10)	2.2846297<H<2.9262175	0.4147565<R<0,54307406
H(20)	3.3370113<H<3.8854821	0.2229035<R<0.3325977
H(30)	2.7737683<H<3.4553927	0.3089214<R<0.4452463

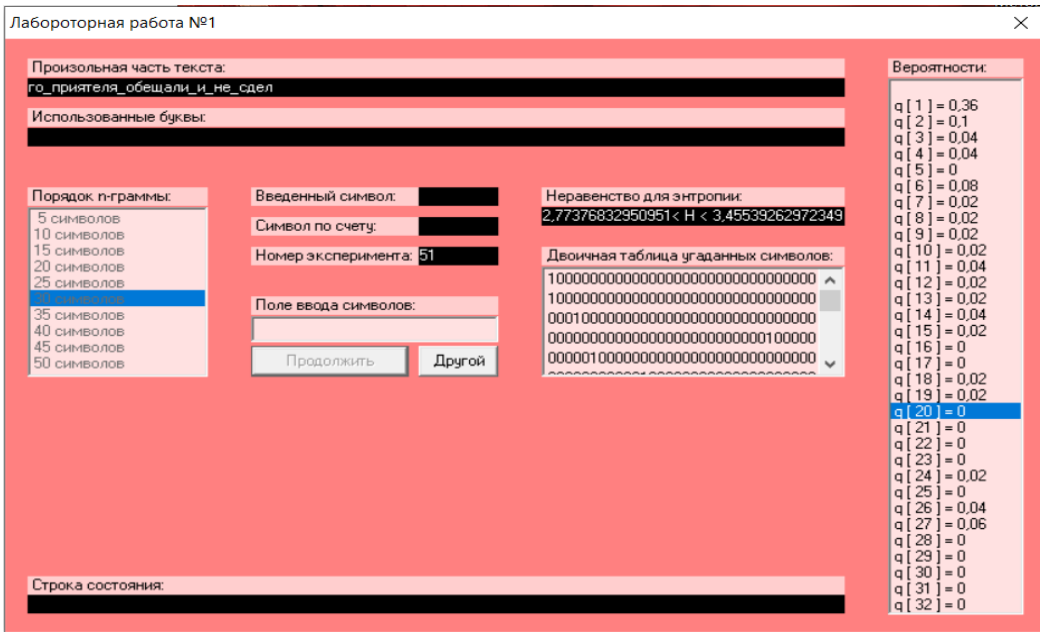
H(10)



H(20)


Лабороторная работа №1
Произольная часть текста:
коны_добра_и_зла_за
Использованные буквы:
Порядок n-граммы:
5 символов
10 символов
15 символов
20 символов
25 символов
30 символов
35 символов
40 символов
45 символов
50 символов
Введенный символ:
Символ по счету:
Номер эксперимента: 53
Поле ввода символов:
Продолжить
Другой
Неравенство для энтропии:
3,33701133943228< H < 3,88548209419984
Двоичная таблица угаданных символов:
Вероятности:
q [1] = 0,2884615
q [2] = 0,1153846
q [3] = 0
q [4] = 0,0384615
q [5] = 0
q [6] = 0,0576923
q [7] = 0,0384615
q [8] = 0,0384615
q [9] = 0,0192307
q [10] = 0,019230
q [11] = 0
q [12] = 0
q [13] = 0,019230
q [14] = 0,019230
q [15] = 0,019230
q [16] = 0
q [17] = 0,019230
q [18] = 0,019230
q [19] = 0,019230
q [20] = 0,019230
q [21] = 0,019230
q [22] = 0,019230
q [23] = 0,019230
q [24] = 0,019230
q [25] = 0,019230
q [26] = 0,057692
q [27] = 0,057692
q [28] = 0,019230
q [29] = 0
q [30] = 0
q [31] = 0,019230
q [32] = 0
Строка состояния:

H(30)


Лабороторная работа №1
Произольная часть текста:
го_приятеля_обещали_и_не_сдел
Использованные буквы:
Порядок n-граммы:
5 символов
10 символов
15 символов
20 символов
25 символов
30 символов
35 символов
40 символов
45 символов
50 символов
Введенный символ:
Символ по счету:
Номер эксперимента: 51
Поле ввода символов:
Продолжить
Другой
Неравенство для энтропии:
2,77376832950951< H < 3,45539262972349
Двоичная таблица угаданных символов:
Вероятности:
q [1] = 0,36
q [2] = 0,1
q [3] = 0,04
q [4] = 0,04
q [5] = 0
q [6] = 0,08
q [7] = 0,02
q [8] = 0,02
q [9] = 0,02
q [10] = 0,02
q [11] = 0,04
q [12] = 0,02
q [13] = 0,02
q [14] = 0,04
q [15] = 0,02
q [16] = 0
q [17] = 0
q [18] = 0,02
q [19] = 0,02
q [20] = 0
q [21] = 0
q [22] = 0
q [23] = 0
q [24] = 0,02
q [25] = 0
q [26] = 0,04
q [27] = 0,06
q [28] = 0
q [29] = 0
q [30] = 0
q [31] = 0
q [32] = 0
Строка состояния:

Висновок:

При виконанні цього комп'ютерного практикуму ми навчились якісно очищувати текст для подальшої роботи над ним. Також ми навчились визначати експериментальним шляхом значення імовірностей(частот) літер та біграм(а за аналогіює й інших n-грам). Користуючись цими значиннями, ми тепер можемо вираховувати ентропію та надлишковість у різних моделях відкритого тексту.

Також ми навчилися працювати з програмою CoolPinkProgram, за допомогою якої змогли наближено обчислити значення H(10), H(20) та H(30).