

# КРИПТОГРАФІЯ

## Комп'ютерний практикум №1

### Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконав: ст. ФБ-12 Слепий Роман

**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.

**Хід роботи:**

1. Ознайомився з методичними вказівками до виконання комп'ютерного практикуму та рекомендаціями стосовно виконання (лайфхаками)
2. Створив програми (4 файли формату .py) для обрахунку частоти та ентропії символів і біграм (з повторами і без) у тексті з пробілами та без.

Для підрахунку ентропії  $H(1)$  символів користався 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

Далі обчислюю H(2). Для цього треба порахувати частоту біграм. Роблю це аналогічно до частоти символів, але з урахуванням того, що частота біграм – відношення кількості появ деякої біграми до загальної кількості біграм у тексті.

Для підрахунку ентропії H(2) біграм(з повторами і без) користався формулою

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

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

$$H_2 = \frac{1}{2} * H(x_1 x_2, \dots, x_n)$$

Де H(x1, x2, x3 ... , xn) в свою чергу, – ентропія n-грами відкритого тексту (x1, x2, x3 ... , xn).

Усі попередні розрахунки були реалізовані у код(усі 4 фрагменти)

Таблиці з результатами виконання коду:

Таблиця для символів(літер) з пробілом:

	Кількість	Частота
«_»	137530	0.16217848
А	54592	0.06437986
Б	13679	0.01613152
В	29428	0.03470418
Г	14105	0.01663390
Д	24446	0.02882895
Е	56302	0.06639645
Ё	5	5.896455876231032e-06
Ж	6466	0.00762529
З	13069	0.01541215
И	49147	0.05795862
Й	7419	0.00874916
К	22018	0.02596563
Л	38145	0.04498406
М	22475	0.02650456

Таблиця для символів(літер) без пробілу:

	Кількість	Частота
А	54592	0.07580200
Б	13679	0.01899355
В	29428	0.04086133
Г	14105	0.01958506
Д	24446	0.03394373
Е	56302	0.07817638
Ё	5	0.00000694
Ж	6466	0.00897816
З	13069	0.01814655
И	49147	0.06824152
Й	7419	0.01030142
К	22018	0.03057240
Л	38145	0.05296504
М	22475	0.03120696
Н	47645	0.06615597

Н	47645	0.05618732
О	80993	0.09551433
П	19952	0.02352921
Р	34322	0.04047563
С	37102	0.04375406
Т	39123	0.04613740
У	20185	0.02380399
Ф	2362	0.00278548
Х	7419	0.00874916
Ц	2171	0.00256024
Ч	9415	0.01110302
Ш	5693	0.00671370
Щ	2092	0.00246707
Ъ	188	0.00022170
Ы	14363	0.01693815
Ь	15212	0.01793937
Э	3255	0.00383859
Ю	3433	0.00404850
Я	14224	0.01677423

Для біграм без повторів,з пробілами:

	Кількість	Частота
о_	18602	0.02193720
и_	17113	0.02018123
е_	14355	0.01692874
_п	13947	0.01644759
_н	13882	0.01637094
_с	12943	0.01526358
а_	12878	0.01518693
_в	12618	0.01488031
_о	9574	0.01129055
_и	9170	0.01129055
То	9032	0.01065137
ь_	8695	0.01025395
я_	8431	0.00994262
но	8341	0.00983648
на	8277	0.00976100
не	8097	0.00954873
по	8027	0.00946618
ст.	7911	0.00932938
ро	7902	0.00931877

О	80993	0.11246029
П	19952	0.02770372
Р	34322	0.04765674
С	37102	0.05151682
Т	39123	0.05432301
У	20185	0.02802725
Ф	2362	0.00327968
Х	7419	0.01030142
Ц	2171	0.00301447
Ч	9415	0.01307290
Ш	5693	0.00790484
Щ	2092	0.00290478
Ъ	188	0.00026104
Ы	14363	0.01994329
Ь	15212	0.02112215
Э	3255	0.00451963
Ю	3433	0.00476678
Я	14224	0.01975029
-	-	-

Для біграм без повторів,без пробілів:

	Кількість	Частота
то	9316	0.012935457
но	8541	0.011859354
на	8352	0.011596924
не	8155	0.011323385
ст	8125	0.011281729
по	8030	0.01114982
ро	7975	0.011073451
ал	7959	0.011051235
ли	7653	0.010626348
он	7131	0.00990154
ра	6850	0.0095113657
ни	6819	0.0094683216
ос	6608	0.0091753438
ко	6508	0.0090364917
ов	6410	0.0089004167
го	6399	0.008885143
от	6099	0.0084685868
од	6078	0.0084394279
ор	6024	0.0083644478

ал	7712	0.00909470
ли	7212	0.00850506
м_	7016	0.00827392
ра	6824	0.00804749
л_	6729	0.00793546
_т	6668	0.00786352
ни	6568	0.00774559
_к	6341	0.00747789
го	6335	0.00747082
ко	6323	0.00745667
й_	5850	0.00689886

Для біграм з повторами, з пробілами:

Біграма	Частота
о_	0.02193717
и_	0.02018121
е_	0.01692872
_п	0.01644757
_н	0.01637092
_с	0.01526357
а_	0.01518691
_в	0.01488030
_о	0.01129053
_и	0.01081410
то	0.01065136
ь	0.01025394
я	0.00994260
но	0.00983647
на	0.00976099
не	0.00954872
по	0.00946617
ст	0.00932937
ро	0.00931876
ал	0.00909469
ли	0.00850505
м_	0.00827391
ра	0.00804748
л_	0.00793545
_т	0.00786351
ни	0.00774558
_к	0.00747789
го	0.00747081
ко	0.00745666
й_	0.00689885

ер	5913	0.008210322
ло	5507	0.00764658
ен	5497	0.00763270
ил	5357	0.00743830
ол	5336	0.00740915
ре	5229	0.00726057
ом	5208	0.00723141
во	5192	0.00720920
пр	5142	0.00713977
ка	5096	0.00707590
ес	4927	0.00684124

Для біграм з повторами, без пробілів:

Біграма	Частота
то	0.01293544
но	0.01185934
на	0.01159691
не	0.01132337
ст	0.01128171
по	0.01114980
ро	0.01107344
ал	0.01105122
ли	0.01062633
он	0.00990153
ра	0.00951135
ни	0.00946831
ос	0.00917533
ко	0.00903648
ов	0.00890040
го	0.00888513
от	0.00846858
од	0.00843942
ор	0.00836444
ер	0.00821031
ло	0.00764657
ен	0.00763269
ил	0.00743829
ол	0.00740914
ре	0.00726056
ом	0.00723140
во	0.00720919
пр	0.00713976
ка	0.00707589
ес	0.00684123

3. За допомогою CoolPinkProgram оцінив значення  $H(10)$ ,  $H(20)$ ,  $H(30)$  та прорахував надлишковість.

### 3.1 Для $H(10)$

Лабораторная работа №1

Произвольная часть текста:  
вовашие\_

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: 60

Поле ввода символов:

Продолжить Другой

Неравенство для энтропии:  
 $1,70468432676532 < H < 2,60544774202607$

Двоичная таблица угаданных символов:

00010000000000000000000000000000
10000000000000000000000000000000
00010000000000000000000000000000
10000000000000000000000000000000
10000000000000000000000000000000
~~~~~

Вероятности:

$q[1] = 0,5254237$
$q[2] = 0,1016949$
$q[3] = 0,1016949$
$q[4] = 0,0508474$
$q[5] = 0,0169491$
$q[6] = 0,0338983$
$q[7] = 0,0338983$
$q[8] = 0,0169491$
$q[9] = 0$
$q[10] = 0$
$q[11] = 0,016949$
$q[12] = 0$
$q[13] = 0,016949$
$q[14] = 0$
$q[15] = 0,016949$
$q[16] = 0$
$q[17] = 0$
$q[18] = 0$
$q[19] = 0,016949$
$q[20] = 0,016949$
$q[21] = 0,016949$
$q[22] = 0,016949$
$q[23] = 0$
$q[24] = 0$
$q[25] = 0$
$q[26] = 0$
$q[27] = 0$
$q[28] = 0$
$q[29] = 0$
$q[30] = 0$
$q[31] = 0$
$q[32] = 0$

Строка состояния:

### 3.2 Для $H(20)$

[illegible]

### 3.3 Для $H(30)$

[illegible]

Користуючись формулою  $R = 1 - \frac{H_{\infty}}{H_0}$  розрахуємо надлишковість для даного тексту.

$H_0 = \log_2 34 \approx 5,08746$  - для словника, до якого входять і пробіли

$H_0 = \log_2 33 \approx 5,04439$  – для словника, до якого не входять пробіли

	Ентропія	Надлишковість
H1(з пробілами)	4.38958257	0,137176
H1(без пробілів)	4.51882074	0,104189
H2(з пробілами, без повторів)	3.98739756	0,216230
H2(з пробілами з повторами)	3.98739371	0,216231
H2( без пробілів, без повторів)	4.18509826	0,170346
H2(без пробілів, з повторами)	4.18509345	0,170358

Для CoolPinkProgram:

	Ентропія	Надлишковість
H(10)	1,70468432<H<2,60544774	0,4878690<R<,664924
H(20)	1,47186376<H<2,26933681	0,5539350<R<,710688
H(30)	1,31257758<H<1,98183722	0,6104470<R<,741997

4.

#### **Висновки:**

*Під час виконання комп'ютерного практикуму я навчився експериментально визначати значення частоти літер та біграм(n-грам) в тексті. З використанням отриманих даних зміг розрахувати ентропію та надлишковість для ВТ. А за допомогою CoolPinkProgram.exe розрахувати ентропію(експериментально) та надлишковість до наданого тексту.*