# 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 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 Н1 та Н2 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 Н(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 Н(2) біграм(з повторами і без) користався формулою

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

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

$$H_{2=\frac{1}{2}*H(x_1x_2,...,x_n)}$$

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

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

Таблиці з результатами виконання коду:

Таблиця для символів(літер) з пробілом:

Таблиця для символів(літер) без пробілу:

Кількість	Частота
137530	0.16217848
54592	0.06437986
13679	0.01613152
29428	0.03470418
14105	0.01663390
24446	0.02882895
56302	0.06639645
5	5.896455876231032e-
	06
6466	0.00762529
13069	0.01541215
49147	0.05795862
7419	0.00874916
22018	0.02596563
38145	0.04498406
22475	0.02650456
	137530 54592 13679 29428 14105 24446 56302 5 6466 13069 49147 7419 22018 38145

	Кількість	Частота
A	54592	0.07580200
Б	13679	0.01899355
В	29428	0.04086133
Γ	14105	0.01958506
Д	24446	0.03394373
Е	56302	0.07817638
Ë	5	0.00000694
Ж	6466	0.00897816
3	13069	0.01814655
И	49147	0.06824152
Й	7419	0.01030142
К	22018	0.03057240
Л	38145	0.05296504
M	22475	0.03120696
Н	47645	0.06615597

H 47645	0.05618732	
O 80993	0.09551433	
П 19952	0.02352921	
P 34322	0.04047563	
C 37102	0.04375406	
T 39123	0.04613740	
У 20185	0.02380399	
Φ 2362	0.00278548	
X 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:		

П бі	£		.:-	
Для біграм	oe3	повтор	ыв,ое	з прооілів:

8541

8352

8155

8125

8030

7975

7959

7653

7131

6850

6819

6608

6508

6410

6399

6099

6078

6024

НО

на

не

ст

ПО

po

ал

ЛИ

OH

pa

НИ

oc

ко

OB

ГО

ОТ

ΟД

op

Кількість	Частота
18602	0.02193720
17113	0.02018123
14355	0.01692874
13947	0.01644759
13882	0.01637094
12943	0.01526358
12878	0.01518693
12618	0.01488031
9574	0.01129055
9170	0.01129055
9032	0.01065137
8695	0.01025395
8431	0.00994262
8341	0.00983648
8277	0.00976100
8097	0.00954873
8027	0.00946618
7911	0.00932938
7902	0.00931877
	18602 17113 14355 13947 13882 12943 12878 12618 9574 9170 9032 8695 8431 8341 8277 8097 8027 7911

O	80993	0.11246029	
Π	19952	0.02770372	
P	34322	0.04765674	
С	37102	0.05151682	
Т	39123	0.05432301	
У	20185	0.02802725	
Φ	2362	0.00327968	
X	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	
R	14224	0.01975029	
-	-	-	
грам без повторів,без пробілів:			

Кількість Частота 9316 0.012935457 TO

0.011859354

0.011596924

0.011323385

0.011281729

0.01114982

0.011073451

0.011051235

0.010626348

0.00990154

0.0095113657

0.0094683216

0.0091753438

 $0.009036\overline{4917}$ 

0.0089004167

0.008885143

0.0084685868

0.0084394279

0.0083644478

ал	7712	0.00909470
ЛИ	7212	0.00850506
M_	7016	0.00827392
pa	6824	0.00804749
л_	6729	0.00793546
_T	6668	0.00786352
ни	6568	0.00774559
_к	6341	0.00747789
ГО	6335	0.00747082
ко	6323	0.00745667
й_	5850	0.00689886

TT	<b>~</b> .		<b>~</b> .
Ππα	OITHOM	з повторами, з	$\Pi h \cap \cap 1 \Pi a M H$
$\Delta IJIJI$	On Dam	3 HODIODAMIN.3	проотлами.
r 1	1	1 ,	1

Linearo	Постото
Біграма	Частота
0	0.02193717
И	0.02018121
e_	0.01692872
_П	0.01644757
_H	0.01637092
_c	0.01526357
a_	0.01518691
_B	0.01488030
_0	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
po	0.00931876
ал	0.00909469
ЛИ	0.00850505
M_	0.00827391
pa	0.00804748
Л_	0.00793545
_T	0.00786351
НИ	0.00774558
_K	0.00747789
ГО	0.00747081
ко	0.00745666
й_	0.00689885

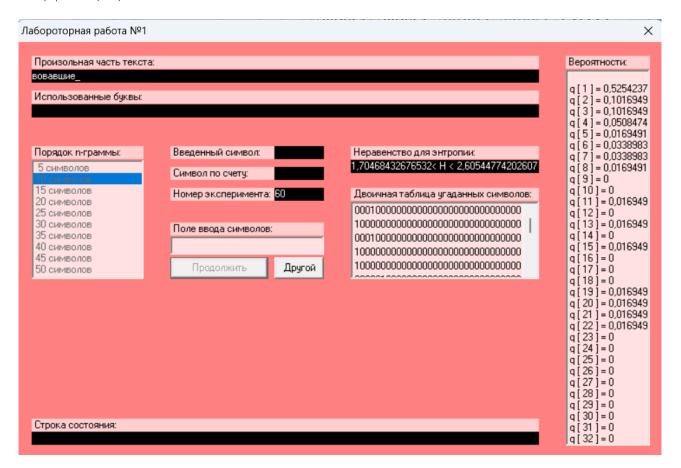
ер	5913	0.008210322
ЛО	5507	0.00764658
ен	5497	0.00763270
ИЛ	5357	0.00743830
ОЛ	5336	0.00740915
pe	5229	0.00726057
OM	5208	0.00723141
во	5192	0.00720920
пр	5142	0.00713977
ка	5096	0.00707590
ec	4927	0.00684124

Для біграм з повторами, без пробілів:

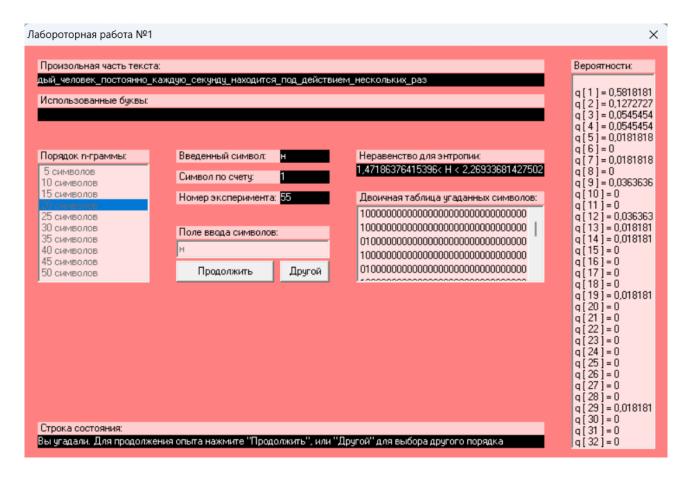
Біграма	Частота
то	0.01293544
НО	0.01185934
на	0.01159691
не	0.01132337
ст	0.01128171
ПО	0.01114980
po	0.01107344
ал	0.01105122
ЛИ	0.01062633
ОН	0.00990153
pa	0.00951135
ни	0.00946831
oc	0.00917533
ко	0.00903648
OB	0.00890040
ГО	0.00888513
ОТ	0.00846858
од	0.00843942
op	0.00836444
ер	0.00821031
ЛО	0.00764657
ен	0.00763269
ИЛ	0.00743829
ОЛ	0.00740914
pe	0.00726056
OM	0.00723140
ВО	0.00720919
пр	0.00713976
ка	0.00707589
ec	0.00684123

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

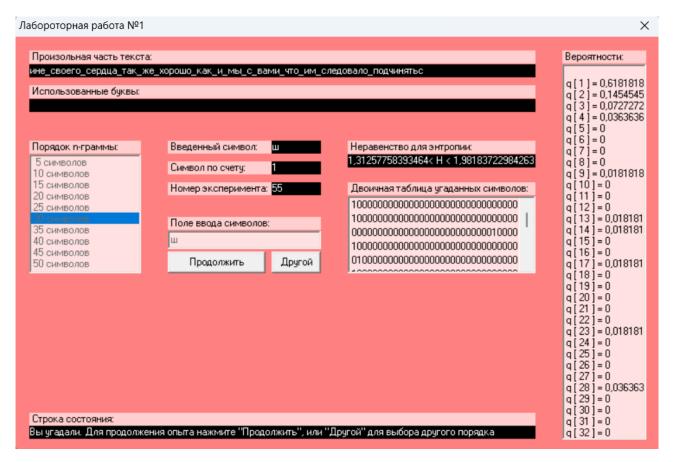
#### 3.1 Для H(10)



# 3.2 Для Н(20)



# 3.3 Для Н(30)



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
Н1(з пробілами)	4.38958257	0,137176
Н1(без пробілів)	4.51882074	0,104189
Н2(з пробілами, без	3.98739756	0,216230
повторів)		
Н2(з пробілами з	3.98739371	0,216231
повторами)		
Н2( без пробілів, без	4.18509826	0,170346
повторів)		
Н2(без пробілів, з	4.18509345	0,170358
повторами)		

#### Для CoolPinkProgram:

	Ентропія	Надлишковість
H(10)	1,70468432 <h<2,60544774< td=""><td>0,4878690<r<,664924< td=""></r<,664924<></td></h<2,60544774<>	0,4878690 <r<,664924< td=""></r<,664924<>
H(20)	1,47186376 <h<2,26933681< td=""><td>0,5539350<r<,710688< td=""></r<,710688<></td></h<2,26933681<>	0,5539350 <r<,710688< td=""></r<,710688<>
H(30)	1,31257758 <h<1,98183722< td=""><td>0,6104470<r<,741997< td=""></r<,741997<></td></h<1,98183722<>	0,6104470 <r<,741997< td=""></r<,741997<>

#### 4.

#### Висновки:

Під час виконання комп'ютерного практикуму я навчився експериментально визначати значення частоти літер та біграм(п-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.