

**Міністерство освіти і науки України Національний
технічний університет України "Київський
політехнічний інститут імені Ігоря Сікорського"
Фізико-технічний інститут**

**КРИПТОГРАФІЯ
КОМП'ЮТЕРНИЙ ПРАКТИКУМ №1**

**Експериментальна оцінка ентропії на символ джерела
відкритого тексту**

**Виконали: студенти групи ФБ-01
Курило А. В. і Шевченко Д. М.**

Київ – 2022

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.

Завдання : 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. За допомогою програми CoolPinkProgram оцінити значення $(10) H$, $(20) H$, $(30) H$. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

Хід роботи:

Таблиці частот букв

З пробілом

1		частота
2	'	0,159721
3	о	0,095254
4	а	0,071039
5	е	0,068993
6	и	0,054999
7	н	0,054873
8	т	0,048573
9	с	0,044328
10	л	0,042227
11	в	0,038323
12	р	0,037193
13	к	0,028984
14	д	0,025689
15	м	0,025366
16	у	0,023369
17	п	0,021265
18	я	0,0191
19	г	0,016946
20	ь	0,01684
21	ы	0,01599
22	б	0,014906
23	з	0,014605
24	ч	0,011957
25	й	0,009731
26	ж	0,008747
27	ш	0,008181
28	х	0,007249
29	ю	0,005505
30	ц	0,00297
31	э	0,002613
32	щ	0,002422
33	ф	0,001616
34	ъ	0,000378
35	ё	4,5E-05

Без пробіла

1		частота
2	о	0,11336
3	а	0,084542
4	е	0,082108
5	и	0,065453
6	н	0,065304
7	т	0,057806
8	с	0,052754
9	л	0,050253
10	в	0,045607
11	р	0,044263
12	к	0,034493
13	д	0,030572
14	м	0,030188
15	у	0,027811
16	п	0,025308
17	я	0,022731
18	г	0,020167
19	ь	0,020041
20	ы	0,019029
21	б	0,017739
22	з	0,017381
23	ч	0,01423
24	й	0,011581
25	ж	0,010409
26	ш	0,009736
27	х	0,008627
28	ю	0,006551
29	ц	0,003535
30	э	0,003109
31	щ	0,002883
32	ф	0,001924
33	ъ	0,00045
34	ё	5,36E-05

Таблиці частот біграм

1	h2_wp_first		h2_wp_second		h2_wop_first		h2_wop_second	
2	'np' : 0.005985701992934504		'np' : 0.006060500437872246		'np' : 0.007123474713954855		'np' : 0.007219399218777006	
3	'pi' : 0.005073650348736307		'ия' : 0.0014624787780822793		'pi' : 0.006150402710083307		'ия' : 0.0019703411801306647	
4	'ия' : 0.0014682869622362693		'тн' : 0.0009120523064902398		'ия' : 0.0019461440077430951		'тн' : 0.0016471360918109855	
5	'ят' : 0.001408742189287024		'or' : 0.0045297629680621935		'ят' : 0.002245151923675205		'or' : 0.006026824293961077	
6	'тн' : 0.0009585256133293153		'o ' : 0.020805830163661106		'тн' : 0.0016384942445297106		'оч' : 0.0026962563517577516	
7	'но' : 0.008222621757369573		'чт' : 0.003539285781714513		'но' : 0.010095405993985274		'че' : 0.00560683051609112	
8	'or' : 0.004541378171031468		'ен' : 0.006828773797957177		'or' : 0.006025095924504822		'ни' : 0.00865740260638114	
9	'ro' : 0.00894478674620493		'то' : 0.013743234038721559		'ro' : 0.010731445953887102		'ят' : 0.0021587334508624562	
10	'o ' : 0.020849384401447953		'м ' : 0.007213636634294619		'оч' : 0.002780946455114245		'ом' : 0.006799405440907048	
11	'ч' : 0.00522614305994779		'не' : 0.0024398851511203867		'чт' : 0.004261294894396626		'не' : 0.002957240139652252	
12	'чт' : 0.003574864844257742		'рв' : 0.00029627176835033265		'те' : 0.005528189705831518		'рв' : 0.0005115973590514709	
13	'те' : 0.0045181411864659085		'ый' : 0.001426170963333464		'ен' : 0.009716893083065435		'ый' : 0.001731826195167479	
14	'ян' : 0.006917359842859892		'ча' : 0.002094234754711665		'ни' : 0.008691969995506239		'ча' : 0.0024681115835320957	
15	'ни' : 0.007014664715728171		'ст' : 0.010861845860255579		'то' : 0.016653703895744755		'ст' : 0.012992153202668602	
16	'то' : 0.013741045591689873		'ь ' : 0.01010664331348022		'ом' : 0.006711258598638045		'п' : 0.0009678868955027827	
17	'ом' : 0.004925514572130868		'ая' : 0.0025894733478855057		'мн' : 0.001039614227937364		'ер' : 0.008183829375367279	
18	'м ' : 0.0071802282307577795		'п ' : 0.014655286345211798		'не' : 0.002910574164333368		'ва' : 0.006789035224169519	
19	'п ' : 0.0148070422586349		'ом' : 0.005001764559796793		'ер' : 0.008296173390023852		'яп' : 0.0011130699298282	
20	'не' : 0.0024442403139897538		'ес' : 0.0035857597845929965		'рв' : 0.0004891285561201562		'ес' : 0.006308548515330637	
21	'ер' : 0.0065484727128816405		'ть' : 0.0055376679054893055		'вы' : 0.0030868678488713748		'ть' : 0.006560890455943862	
22	'рв' : 0.00029627155321087926		'я ' : 0.010777611730038326		'ый' : 0.0017171350547893118		'ям' : 0.0007552974523834215	
23	'вы' : 0.002593828402130541		'мо' : 0.003581402846823139		'йн' : 0.00037073524836669084		'ой' : 0.004277714404231048	
24	'ый' : 0.001442871510367689		'й ' : 0.007732112228907701		'ча' : 0.0024534204431539285		'ве' : 0.005319921186352795	
25	'йч' : 7.479404407039354e-05		'ве' : 0.00434386695654826		'ас' : 0.006374226554668326		'рн' : 0.000895295378340074	
26	'ча' : 0.0020586516013550065		'рн' : 0.0006651591661982958		'ст' : 0.013051781948909399		'ра' : 0.009716893083065435	
27	'ас' : 0.003715739063186444		'р ' : 0.004060666001507501		'ть' : 0.006556569532303225		'бн' : 0.000324933457759342	
28	'ст' : 0.010712976040490445		'аб' : 0.0006230421010896701		'ьн' : 0.0010197379791904318		'ук' : 0.0015451622938919423	
29	'н ' : 0.0055093438093405415		'н ' : 0.0148513485448554		'яб' : 0.006727678108472467		'ня' : 0.003752290089529538	
30	'ь ' : 0.01008194667838198		'у ' : 0.006282704264134995		'ая' : 0.0033357530505720902		'зь' : 0.0010231947181029418	
31	'ва' : 0.005573245516895829		'кн' : 0.0018865540543484417		'ян' : 0.0011381312869438972		'ре' : 0.0006014725707767292	
32	'ая' : 0.0026112561405547104		'яз' : 0.0012722258287984872		'по' : 0.009792941339140654		'ну' : 0.0031784714300528883	
33	'я ' : 0.010800840888378967		'ге' : 0.0004690969665546934		'ме' : 0.003460195651422448		'ий' : 0.0010283798264717067	
34	'по' : 0.00822734484774329		'ну' : 0.002412291211911287		'еc' : 0.0063552144906495215		'лу' : 0.0016212105499671609	

*на скріншоті зазначено лише фрагмент з таблиці

Топ 10 букв і біграм для кожного з випадків

h1_wp		h1_wop		h2_wp_first		h2_wp_second		h2_wop_first		h2_wop_second	
	кількість		кількість		кількість		кількість		кількість		кількість
" "	219955	о	131176	"о "	28712	"о "	14326	то	19271	то	9702
о	131176	а	97829	"и "	24570	"и "	12318	ст	15103	ст	7517
а	97829	е	95012	"а "	23347	"а "	11713	на	14836	на	7404
е	95012	и	75740	"е "	22745	"е "	11443	ов	13442	ов	6793
и	75740	н	75567	"с "	21920	"с "	10974	ал	12566	не	6342
н	75567	т	66891	"н "	20576	"н "	10226	не	12560	го	6228
т	66891	с	61045	"п "	20391	"п "	10091	го	12418	ал	6200
с	61045	л	58151	"в "	19972	"в "	9934	он	12046	он	5972
л	58151	в	52775	"то "	18923	"то "	9463	ос	11789	ос	5895
в	52775	р	51219	"о "	16198	"о "	8119	ко	11759	ко	5892

Ентропія та надлишковість

Ентропія h1 з пробілами: 4.388894879952538	Надлишковість: 0.13731165869825102
Ентропія h1 без пробів: 4.4690563333633255	Надлишковість: 0.11405488397252739
Ентропія h2 з пробілами для першого випадку: 3.982418981688493	Надлишковість: 0.21720922472433446
Ентропія h2 з пробілами для другого випадку: 3.9816166933944723	Надлишковість: 0.21736692382092848
Ентропія h2 без пробілів для першого випадку: 4.148115315384743	Надлишковість: 0.17767818746242203
Ентропія h2 без пробілів для другого випадку: 4.148140197144114	Надлишковість: 0.17767325490584884

CoolPinkProgram

$$2,46184846352643 < H_{10} < 3,206890062718$$

$$0,3586219874564 < R < 0,5076303072947$$

Лабораторная работа №1

×

Произвольная часть текста:
ащается_к

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:
2,46184846352643 < H < 3,206890062718

Двоичная таблица угаданных символов:
00000000000010000000000000000000
10000000000000000000000000000000
00010000000000000000000000000000
10000000000000000000000000000000
10000000000000000000000000000000
00000000000000000000100000000000

Вероятности:
q[1] = 0.42
q[2] = 0.1
q[3] = 0.09
q[4] = 0.04
q[5] = 0
q[6] = 0.02
q[7] = 0.04
q[8] = 0
q[9] = 0.02
q[10] = 0.02
q[11] = 0
q[12] = 0.04
q[13] = 0
q[14] = 0.04
q[15] = 0
q[16] = 0
q[17] = 0
q[18] = 0.02
q[19] = 0.02
q[20] = 0
q[21] = 0.02
q[22] = 0
q[23] = 0.04
q[24] = 0.02
q[25] = 0
q[26] = 0
q[27] = 0.02
q[28] = 0
q[29] = 0
q[30] = 0
q[31] = 0.02
q[32] = 0.02

Строка состояния:

$$1,66900373601988 < H_{20} < 2,331346433249$$

$$0,5337307133502 < R < 0,666199252796024$$

Лабораторная работа №1

×

Произвольная часть текста:
момент_когда_у_вас

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:
1,66900373601988 < H < 2,331346433249

Двоичная таблица угаданных символов:
10000000000000000000000000000000
01000000000000000000000000000000
01000000000000000000000000000000
10000000000000000000000000000000
10000000000000000000000000000000
10000000000000000000000000000000

Вероятности:
q[1] = 0.56
q[2] = 0.12
q[3] = 0.06
q[4] = 0.02
q[5] = 0.04
q[6] = 0
q[7] = 0
q[8] = 0
q[9] = 0
q[10] = 0.06
q[11] = 0
q[12] = 0
q[13] = 0.04
q[14] = 0.02
q[15] = 0
q[16] = 0.04
q[17] = 0
q[18] = 0
q[19] = 0
q[20] = 0
q[21] = 0
q[22] = 0
q[23] = 0
q[24] = 0
q[25] = 0
q[26] = 0.02
q[27] = 0
q[28] = 0
q[29] = 0
q[30] = 0
q[31] = 0.02
q[32] = 0

Строка состояния:

$$0,702303900072 < R < 0,8208938822567$$

×

Произвольная часть текста:		
епредвиденное_освобождающее_e		
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:		
0,895530588716437 < H < 1,48848049964		
Двоичная таблица угаданных символов:		
10000000000000000000000000000000 10000000000000000000000000000000 00100000000000000000000000000000 01000000000000000000000000000000 10000000000000000000000000000000 01000000000000000000000000000000		

Вероятности:

- q[1] = 0,72
- q[2] = 0,14
- q[3] = 0,04
- q[4] = 0
- q[5] = 0
- q[6] = 0
- q[7] = 0
- q[8] = 0,02
- q[9] = 0
- q[10] = 0,02
- q[11] = 0,02
- q[12] = 0
- q[13] = 0
- q[14] = 0
- q[15] = 0
- q[16] = 0,02
- q[17] = 0
- q[18] = 0
- q[19] = 0
- q[20] = 0
- q[21] = 0
- q[22] = 0
- q[23] = 0
- q[24] = 0,02
- 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:		

Висновки : Під час лабораторної роботи ми розглянули та порівняли різні моделі джерела відкритого тексту для наближеного визначення ентропії. Опрацювавши значення ентропії на символ джерела відкритого тексту та провівши практичну частину роботи, мали змогу експериментально оцінити саму ентропію та надлишковість, таким чином отримали детальні знання щодо цих понять.